PPRs and cpRNPs: RNA-binding proteins required for global RNA stabilization in plant organelles

Dissertation

zur Erlangung des akademischen Grades
d o c t o r r e r u m n a t u r a l i u m
(Dr. rer. nat.)

im Fach Biologie
eingereicht an der Lebenswissenschaftlichen Fakultät
der Humboldt-Universität zu Berlin
von

Diplom-Biochemiker Hannes Ruwe

Präsident der Humboldt-Universität zu Berlin
Prof. Dr. Jan-Hendrik Olbertz

Dekan der Lebenswissenschaftlichen Fakultät
Prof. Dr. Richard Lucius

Gutachter/in: 1. Prof. Dr. Christian Schmitz-Linneweber
2. Prof. Dr. Ian David Small
3. Prof. Dr. Wolfgang Schuster

Tag der mündlichen Prüfung: 07.07.2015

Abstract

Chloroplasts and mitochondria are of endosymbiotic origin. Their basic gene expression machineries are retained from their free-living prokaryotic progenitors. On top of this bacterial scaffold, a number of organelle-specific RNA processing steps evolved. These include RNA editing and processing of polycistronic mRNAs into smaller units of mono- and dicistronic mRNAs. In general, regulation of gene expression has shifted from typical prokaryotic transcriptional regulation to regulation on the posttranscriptional level.

In this thesis, a novel class of organelle-specific short (15-50nt) RNAs is described on a transcriptome-wide scale. The small RNAs are found at binding sites of PPR (Pentatricopeptide repeat) and PPR-like proteins, which protect mRNAs against exonucleolytic decay. The small RNAs represent minimal nuclease resistant RNAs, so called PPR footprints. Small RNAs were identified in almost every intergenic region subjected to intergenic processing. This finding suggests that accumulation of processed transcripts in plastids is mostly due to protection by highly specific RNA-binding proteins. Small RNA sequencing identified a number of nuclease insensitive sites missing in mutants of RNA-binding proteins. Analysis of multiple small RNAs representing target sites of single PPR proteins expands the knowledge of target specificity. A catalogue of orphan small RNAs identified in this thesis awaits the assignment of their cognate RNA-binding proteins. In mitochondria, accumulations of small RNAs predicts that at least two thirds of mitochondrial mRNAs are stabilized by RNA-binding proteins binding in their 3'UTR. In sum, small organellar RNAs turned out to be instrumental in elucidating the hitherto enigmatic intercistronic processing of organellar RNAs and allowed novel insights into the function of the dominant family of organellar RNA binding proteins, the PPR proteins.

A chloroplast ribonucleoprotein CP31A is shown to be involved in stabilization of an mRNA for a central component of the NDH-complex by interaction with its 3'UTR. In addition, CP31A represents the first factor described that influences the accumulation of chloroplast antisense transcripts.

Finally, ten novel plastid C to U RNA-editing sites were identified in the model plant *Arabidopsis thaliana*, using a novel RNA-Seq based approach.

Keywords: PPR protein, chloroplast, plastid, mitochondria, ribonucleoprotein, RNA processing, RNA editing, RNA stability, small RNA, non-coding RNA

Zusammenfassung

Die Genexpressionsmaschinerie in Chloroplasten und Mitochondrien und die ihrer prokaryotischen Vorläufer sind konserviert. Innerhalb eines bakteriellen Grundgerüsts entwickelte sich darüber hinaus ein komplexer RNA-Metabolismus. Organellen-spezifische Schritte beinhalten RNA-Edierung und die Prozessierung von zunächst polycistronischen Vorläufertranskripten zu mono- und dicistronischen Einheiten. Grundsätzlich kann man evolutionär von einer Verschiebung hin zu posttranskriptioneller Kontrolle der Genexpression sprechen.

In der vorliegenden Arbeit wird eine neue Klasse kleiner RNAs (15-50nt) mit plastidärem und mitochondrialen Ursprung beschrieben. Diese kurzen RNAs überlappen mit Bindestellen von RNA-bindenden Proteinen, die mRNAs gegen exonukleolytischen Verdau beschützen. Diese stabilisierende Funktion wird vermutlich hauptsächlich von PPR (Pentatricopeptid repeat) Proteinen und verwandten Proteine bewerkstelligt. Die kleinen RNAs repräsentieren dabei minimale nuklease-resistente Bereiche, sogenannte RNA-Bindeprotein *footprints*. Solche *footprints* finden sich in fast jedem intergenischen Bereich, der Prozessierung aufweist. Durch transkriptomweite Untersuchungen von kleinen RNAs in Mutanten von RNA-Bindeproteinen konnte für diese eine Reihe von Bindestellen identifiziert werden. Nuklease-resistente kleine RNAs fehlen in entsprechenden Mutanten. Der Vergleich neu identifizierter Ziele einzelner RNA-Bindeproteine führte dabei zu neuen Erkenntnissen über den Mechanismus der RNA-Erkennung durch PPR Proteine. Im Gegensatz zu Plastiden befinden sich kleine RNAs in Mitochondrien überwiegend an den 3‘ Enden von Transkripten, deren Stabilität vermutlich maßgeblich von diesen RNA-Bindeproteinen beeinflusst wird.

Für das chloroplastidäre Ribonukleoprotein CP31A konnte gezeigt werden, dass es an der Stabilisierung der *ndhF* mRNA beteiligt ist. Die Interaktion mit der *ndhF* mRNA, die eine zentrale Komponente des NDH-Komplexes kodiert, wird dabei über die 3‘ untranslatierte Region vermittelt. Zusätzlich konnte gezeigt werden, dass CP31A die Stabilität einiger antisense Transkripte beeinflusst.

Weiterhin wurden zehn neue Cytidin Desaminierungungen durch die Analyse von RNA-Seq Datensätzen in der Modellpflanze *Arabidopsis thaliana* identifiziert.

Schlagworte: PPR Protein, Chloroplast, Plastid, Mitochondrium, Ribonukleoprotein, RNA Prozessierung, RNA Edierung, RNA Stabilität, kleine RNA, nichtkodierende RNA

Table of contents

1 Introduction

1.1 Endosymbiotic origin of plastids and mitochondria

The evolution of eukaryotic cells is closely connected to a DNA-containing organelle, the mitochondrion. Of proteobacterial origin, mitochondria are believed to have evolved only once by endosymbiosis (reviewed in Zimorski et al. 2014). Plants harbor a second endosymbiont that is of cyanobacterial origin, and is found in various differentiated forms inside a plant, the plastid (reviewed in Pyke 1999). The best studied form is the chloroplast, which can perform photosynthesis, the basis for the photoautotrophic life style of plants.

1.2 Organellar gene expression in plants

Once free-living, mitochondria and plastids contain genomic information stored in small genomes of circular and linear nature (reviewed in Backert et al. 1997, reviewed in Bendich 2004). The gene expression systems in organelles retained many prokaryotic features, with organelle-specific differences believed to display adjustments to the life inside a host cell. The following sections focus on gene expression in plastids, but where supportive for the thesis, mitochondrial features are described as well.

1.2.1 Organellar genomes

The integration of mitochondria and plastids into the host cell was accompanied by a massive loss of genetic information in the organelles. Organellar genomes of present day plants contain about 100 genes in plastids and even fewer genes in mitochondria. Many of the endosymbiont genes have been lost entirely or transferred to the nuclear genome (reviewed in Timmis et al. 2004). Transfer events are still happening and recent transfer events are evident in so called nuclear plastid DNA (NUPTs) and nuclear mitochondrial DNA (NUMTs), (Michalovova et al. 2013). Many gene products are posttranslationally reimported into the two organelles and in general imported proteins represent the majority of the organellar proteomes (reviewed in Leister 2003).

Genes retained in the organellar genomes encode subunits of photosynthetic complexes in plastids or oxidative phosphorylation in mitochondria. In addition, genes encoding ribosomal RNAs and transfer RNAs are present, as well as genes encoding ribosomal proteins and a plastid encoded multisubunit plastid RNA polymerase. A few additional gene

products are required for protein import into the chloroplast, proteolysis and fatty acid synthesis.

Organellar genomes are found in multiple copies per chloroplast and are organized in so called nucleoids, which contain several copies of plastid chromosomes (reviewed in Powikrowska et al. 2014). Plastid genomes of land plants share a characteristic architecture. Two single copy regions are separated by two inverted repeat sequences in which the ribosomal RNA operon resides. Plastid chromosomes are gene-dense with the about 100 genes dispersed in a genome of about 150kb.

1.2.2 Transcription

Organellar genes are transcribed by two types of RNA polymerases. Plastids encode a multisubunit polymerase of eubacterial origin, supported by nuclear-encoded sigma factors for DNA recognition. In addition, two nuclear-encoded phage-type polymerases are imported into plastids of dicot plants. One of this single-subunit polymerases is dually targeted and also resides in mitochondria. A third phage-type polymerase is imported into mitochondria alone (reviewed in Liere et al. 2011).

For plastids, a share of labor between the plastid-encoded plastid RNA polymerase (PEP) and the nuclear-encoded plastid RNA polymerase (NEP) has been proposed. The NEP enzymes transcribe housekeeping genes, like ribosomal protein genes and the genes for the PEP. The PEP transcribes photosynthetic genes and is more active in later stages of chloroplast development from proplastids (Hajdukiewicz et al. 1997). Single gene and genome-wide promoter analysis showed that many genes can be transcribed by both RNA polymerases showing that if a division of labor exists, it is not absolute (Hajdukiewicz et al. 1997, Zhelyazkova et al. 2012b). Both polymerases are essential for the development of photosynthetically active chloroplasts (Allison et al. 1996, Hricova et al. 2006). Genome-wide investigations identified a number of transcriptions initiation sites, with several found inside of open reading frames and antisense to these, resulting in a plethora of non-coding transcripts (Hotto et al. 2011, Zhelyazkova et al. 2012b).

Promoter recognition by PEP is modulated by nuclear-encoded sigma factors that share similarity with *E.coli* σ^{70} and promoter elements resemble σ^{70}-recognized sequences. Six sigma factors with partially overlapping functions have been identified in *Arabidopsis* and potentially regulate chloroplast gene expression on a transcriptional level (reviewed in Lerbs-Mache 2011). Circadian oscillation of *psbD* transcription was recently traced back

to circadian control of the sigma factor Sig5 in the nucleus, supporting that sigma factors can play regulatory roles in chloroplast gene expression (Noordally et al. 2013).

Transcriptional activity and steady-state levels of chloroplast and mitochondrial transcripts was shown to not correlate well for many genes (Deng and Gruissem 1987, Deng et al. 1989, Giege et al. 2000, Holec et al. 2006). This lead to the conclusion that gene expression in organelles is predominantly controlled at the post-transcriptional level.

1.2.3 Translation

Plastid ribosomes resemble 70S bacterial ribosomes with a protein composition inherited from their cyanobacterial ancestors (Yamaguchi and Subramanian 2003, Yamaguchi et al. 2000). About 60% of plastid genes exhibit Shine-Dalgarno or Shine-Dalgarno-like sequences close to start codons, which in bacteria interact with the 16S rRNA to recruit ribosomes for translation (Scharff et al. 2011). Shine-Dalgarno free mRNAs display reduced RNA structure around the start codon in bacteria and both plastids and mitochondria in plants (Scharff et al. 2011).

Translation of some plastid genes is rapidly increased by light (Klein et al. 1988) and translational activity can counterbalance reduced transcript levels, artificially induced by inhibition of transcription (Eberhard et al. 2002). Thus translation was proposed to be the rate-limiting step for many plastid genes. Translational activation executed by nuclear-encoded RNA-binding proteins (RBPs) was shown for a number of plastid genes in the unicellular algae *Chlamydomonas reinhardtii* and higher plants. RBPs Nac2 and RBP40 act in a complex on the *psbD* mRNA. In *Chlamydomonas* Nac2 stabilizes *psbD* and RBP40 is required for efficient translation (Schwarz et al. 2007). Similarly a complex of MCA1 and TCA1 binds the 5' UTR of *petA* with MCA1 required for stability and TCA1 required for translation of cytochrome f in *Chlamydomonas* (Loiselay et al. 2008, Raynaud et al. 2007). MCA1 and TCA1 are rate limiting for cytochrome f translation. Furthermore, MCA1 is destabilized by unassembled cytochrome f resulting in a feedback inhibition (Boulouis et al. 2011). In higher plants several RNA-binding proteins, belonging to the family of helical repeat proteins, were implicated in translational activation of a single or a small number of plastid transcripts (Barkan et al. 1994, Cai et al. 2011, Pfalz et al. 2009, Sane et al. 2005). PPR10 and HCF107 were hypothesized, based on *in vitro* experiments, to reduce RNA structure around the Shine-Dalgarno sequence and the start codon, resulting

in the liberation of a ribosome landing pad (Hammani et al. 2012, Pfalz et al. 2009). Similarly in mitochondria of maize and *Arabidopsis*, MPPR6 interacts with the 5' UTR of *rps3* transcripts and is required for translation of the downstream open reading frame (Manavski et al. 2012).

1.2.4 RNA processing

Genes in plastids are often arranged in operons. The two polymerase activities transcribe polycistronic messages which undergo a massive amount of post-transcriptional processing. 5' and especially 3' ends are trimmed, introns removed and coding is altered by RNA editing. Initial polycistronic messages are frequently processed into smaller units of mono or dicistronic mRNAs. In mitochondria, poly-cistronic messages are less prominent, accordingly intercistronic processing is less frequent. All other processing steps described for plastids are similarly found in mitochondria of land plants.

1.2.4.1 RNA splicing

Slightly less than 20 introns interrupt plastid genes of land plants. All but one belong to the group II introns. The exception is an intron found in the *trnL*-UAA, a member of the group I introns [group I and II introns can be distinguished based on conserved structure and also the splicing mechanism (reviewed in de Longevialle et al. 2010)]. Bacterial group I and II introns show autocatalytic splicing *in vitro*. In contrast, chloroplast introns lost this ability and require several protein factors for correct intron splicing *in vivo* (reviewed in Germain et al. 2013). The protein factors are believed to guide intron folding to a final catalytic active structure (Ostersetzer et al. 2005). Several atypical RNA-binding motifs are found in chloroplast splicing factors. In addition, proteins containing RNA-recognition motifs (RRMs) and pentatricopeptide-repeat (PPR) proteins are involved in intron splicing (reviewed in Germain et al. 2013). Chloroplast splicing factors often support splicing of more than one intron, but a chloroplast "spliceosome" that targets all plastid introns does not exist (reviewed in Barkan 2011). In bacteria, group I and II introns often contain an open reading frame that encodes a maturase protein that helps self-splicing by stabilizing catalytic active structures. In chloroplasts, a protein with similarities to bacterial maturases is encoded in the *trnK*-UUU intron. MatK was shown to interact with several group IIA

introns *in vivo* including its home intron (Zoschke et al. 2010). In contrast, bacterial maturases usually only guide splicing of the home intron (reviewed in Lambowitz and Zimmerly 2004).

A regulatory role for splicing in organellar gene expression can be envisioned. Unspliced precursor transcripts accumulate to substantial amounts in plastids and splicing status was shown to differ in various tissues and developmental stages (Barkan 1989, Hertel et al. 2013). Investigations on weak alleles for plastid splicing factor RNC1 indicate that the amount of splicing factors can be limiting for intron splicing (Watkins et al. 2007). Evidence for splicing being rate-limiting for plastid gene expression is however lacking.

1.2.4.2 Intergenic and end processing

Termination of transcription and to some extend transcription initiation is relaxed in plastids and mitochondria. This leads to an initial accumulation of transcripts representing large parts of the organellar genomes. Many unwanted transcripts are rapidly degraded by a number of ribonucleases present in the two organelles (reviewed in Germain et al. 2013). Trimming of 3' extensions compensates for inefficient transcription termination by organellar polymerases (Figure 1), (Stern and Gruissem 1987). Transcription initiation on opposite strands results in the generation of antisense RNAs which are rapidly degraded predominantly by a 5'→3' exonucleolytic activity in chloroplasts (Sharwood et al. 2011). Exonucleases can be blocked by stable stem-loop structures or, as recently demonstrated, by RNA-binding proteins in 5' and 3' untranslated regions (UTRs), (Pfalz et al. 2009, Prikryl et al. 2011). By this roadblock mechanism stable RNA structures and RBPs determine which transcripts accumulate and which are subjected to degradation. In mitochondria of land plants 3'→5' exonucleases, namely the PNPase and RNase II, have been shown to trim 3' ends and degrade superfluous transcripts (Giege et al. 2000, Holec et al. 2006). Similar as in plastids, an RBP was found to bind in the 3' UTR of *nad4* and blocks exonucleases thereby stabilizing the mRNA (Haili et al. 2013).

Polycistronic as well as mono- or dicistronic mRNAs accumulate to substantial amounts. Initially it was assumed that separation of individual cistrons was a result of a single endonucleolytic cleavage event. This assumption was based on inaccurate mapping of transcript termini (reviewed in Barkan 2011). Precise transcript mappings in the *petB-petD* intergenic region showed that 3' ends of upstream and 5' ends of downstream processed cistrons overlap by about 30nt (Barkan et al. 1994). Similarly overlapping transcript

ends were identified for processed transcripts in the *atpH-atpI*, *psaJ-rpl33* and *psbH-petB* intergenic regions (Pfalz et al. 2009). Accumulation of processed transcript ends was shown to depend on the presence of specific RNA-binding proteins (Barkan et al. 1994, Pfalz et al. 2009). A model for intercistronic processing was proposed, based on findings for PPR protein PPR10 (Figure 1 and Figure 2). According to this model, processing is indeed initiated by endonucleolytic cleavage, but rather in a stochastic, nonspecific way. These cleavage sites act as entrance sites for exoribonucleases like PNPase and RNase J that degrade mRNAs until they are blocked by the next RBP or stable RNA structure (reviewed in Barkan 2011). Thus RNA decay and RNA processing rely on the same factors.

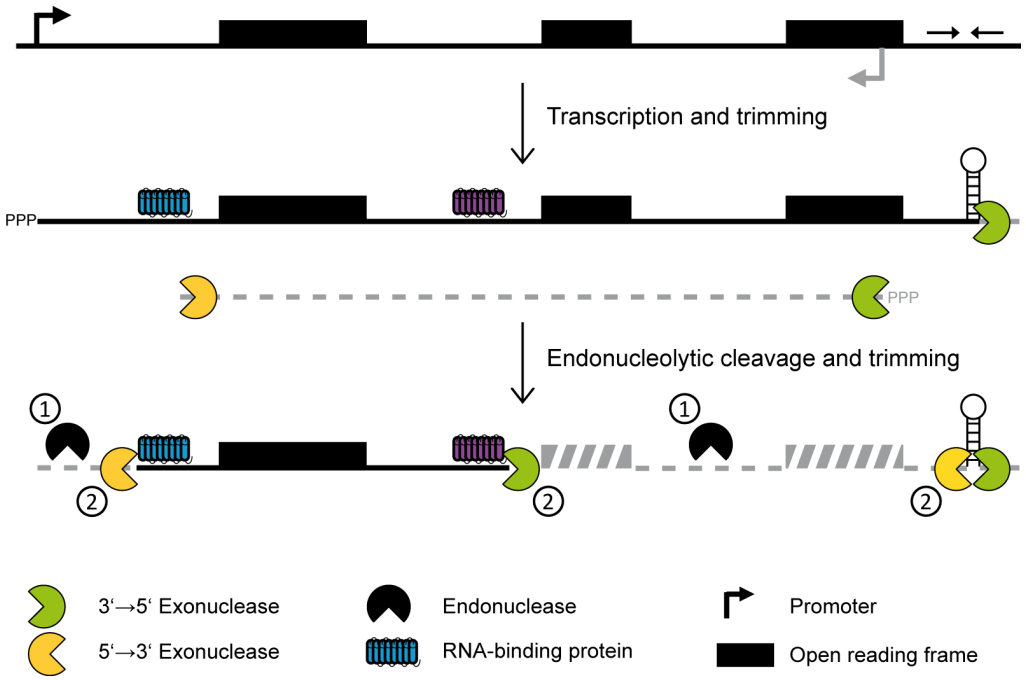


Figure 1: Model for the trimming and processing of plastid transcripts. Plastid transcripts are often initially polycistronic. Primary transcripts are trimmed at their 3' end by exonucleases that are sensitive to stable structures and stably bound RNA-binding proteins. Antisense transcripts that result from relaxed transcription initiation are often rapidly degraded by plastid RNase J. Endonucleolytic cleavage inside of polycistronic mRNAs creates entrance points for exonucleases that degrade RNA until they reach the next stable structure or RNA-binding protein.

1.2.4.3 RNA stability and degradation

The mechanism of protein-mediated protection of RNAs seems to be an organelle-specific feature, not present in cyanobacteria or proteobacteria. This mechanism could explain, at least in part, the longevity of organellar mRNAs. Although the set of ribonucleases in plastids shows strong resemblance of their cyanobacterial counterparts, mRNA half-lives

have been estimated to be an order of magnitude higher in plastids, to be measured in hours rather than minutes (Germain et al. 2012, Klaff and Gruissem 1991). In bacteria, endonucleolytic cleavage is rate-limiting for RNA decay. Similar evidence for the dependence on endonucleases is not as clear in plastids, where exonucleolytic activity might be equally important (reviewed in Germain et al. 2013). Either a reduction of ribonuclease activity or an increase in protective factors, likely RNA-binding proteins, have been speculated to result in the long half-lives found for plastid mRNAs (reviewed in Germain et al. 2013). Chloroplast ribonucleoproteins (cpRNPs), a class of highly abundant RNA-binding proteins, have been shown to bind untranslated mRNAs and likely protects them against endonucleolytic cleavage (Nakamura et al. 2001), (see 1.2.5.3). Helical repeat proteins and potentially other RNA-binding proteins are able to block exonucleases by the roadblock mechanism described above (1.2.4.2).

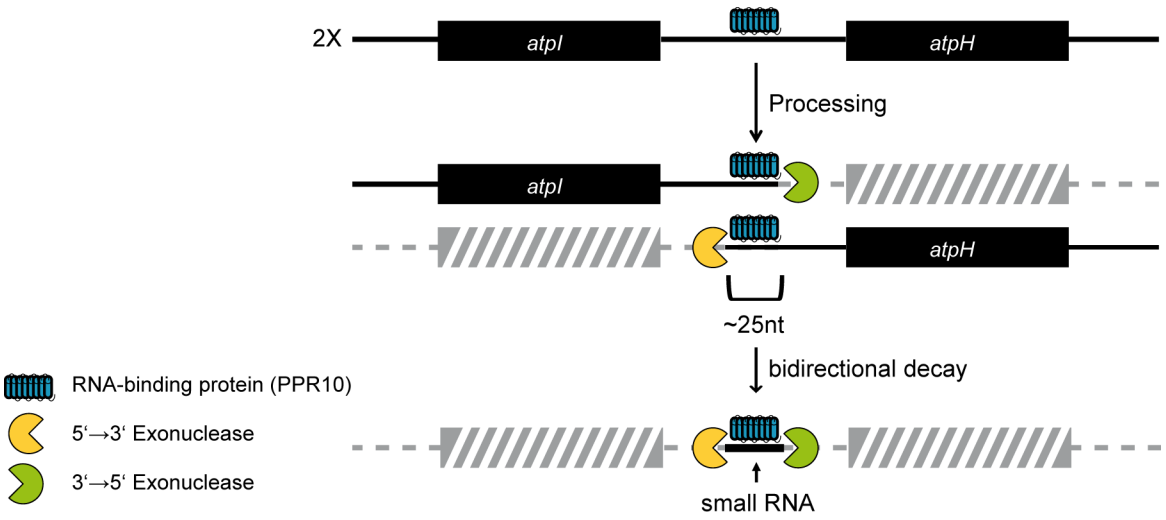


Figure 2: Model for the generation of overlapping transcript ends and RBP footprints. Mapping of transcript termini in the intergenic region between *atpI* and *atpH* revealed overlapping processed transcript termini. PPR protein PPR10 was proposed to block both 5'→3' and 3'→5' exonucleases, thereby stabilizing processed transcript ends. Protection against both types of exonucleases results in the accumulation of a small RNA representing the overlap of up- and downstream processed transcripts.

The roadblock mechanism can explain the finding of overlapping transcript ends, with 5'→3' exonucleases stopped upstream and 3'→5' exonucleases stopped downstream of a bound RNA-binding protein (Figure 2). What happens if a processed mRNA is finally degraded? The model depicted in Figure 2 shows that a bidirectional decay could lead to the accumulation of a small RNA with the sequence representing the overlap of the pro-

cessed up- and downstream cistron. Indeed, a small RNA representing the overlap of processed *atpI* and *atpH* was found in a small RNA database and assumed to represent the footprint of PPR protein PPR10 (Pfalz et al. 2009).

1.2.4.4 RNA editing

In organelles of land plants, a number of recoding events is found at the level of RNA. In seed plants 30-40 C→U changes are observed in plastids and more than 400 C→U changes in mitochondrial transcripts. In ferns, mosses and Lycopodiaceae the reverse reaction, U→C, is also frequently found (reviewed in Takenaka et al. 2013b). RNA editing was identified predominantly in coding regions were RNA editing usually restores codons for conserved amino acids (Gualberto et al. 1989, Maier et al. 1992). RNA-editing efficiency at specific sites has been shown to respond to developmental and environmental cues. In general, RNA editing has the potential to regulate organellar gene expression, but evidence for a true regulatory role of an editing event is missing (reviewed in Takenaka et al. 2013b).

Editing sites in plastids or mitochondria do not share a consensus sequence, and a number of RNA-binding proteins of the PPR protein family have been implicated in recognizing the variable sequences upstream of the Cs to be edited. The first of such trans-factors identified was CRR4, which is required for editing the second base in the *ndhD* open reading frame, resulting in the generation of the AUG start codon (Kotera et al. 2005). CRR4 belongs to the PLS-class of PPR proteins, which almost without exception carry additional C-terminal domains, namely the E domain and for about half the PLS-class proteins an additional DYW domain (see 1.2.5.1). All later identified specific trans-factors in chloroplasts similarly belong to the PLS-class and at least carry an E domain (reviewed in Shikanai 2015, Wagoner et al. 2015, Yap et al. 2015). The enzymatic activity, the cytidine deaminase, has not been identified yet. Potentially it resides in the C-terminal DYW domain. The DYW domain shows similarities with cytidine deaminase domains and was speculated to be recruited from other trans-factors if missing (Boussardon et al. 2012, Hayes et al. 2013, Iyer et al. 2011, Salone et al. 2007). The E domain is essential for the editing reaction. Two particularly degenerated PPR repeats can be predicted in the E domain and are speculated to be required for protein-protein interaction with the editing activity or even RNA-binding (Okuda et al. 2007, reviewed in Takenaka 2014, Wagoner et al. 2015, Yagi et al. 2013).

Even though PPR proteins are required for site recognition and potentially carry the enzymatic activity in their DYW domains, they are not sufficient for efficient editing at most sites. Proteins belonging to the multiple organellar RNA-editing factor family (MORFs) that were identified at the same time in a different group and named RNA-editing factor interacting proteins (RIPs), are required for efficient editing at a large number of editing sites (Bentolila et al. 2012, Bentolila et al. 2013, Takenaka et al. 2012). Furthermore, RIP/MORF proteins were shown to interact with PPR proteins involved in RNA editing (Bentolila et al. 2012, Takenaka et al. 2012). CP31A an RNA-binding protein of the chloroplast ribonucleoprotein family (cpRNPs) is required for the efficient editing at a subset of plastid RNA-editing sites (Tillich et al. 2009).

1.2.5 RNA-binding proteins in plastids and mitochondria of land plants

The different posttranscriptional steps described in the previous sections highlight the complexity of RNA metabolism in plant organelles. Not surprisingly, a large number of RNA-binding proteins is imported into chloroplasts and mitochondria. With the exception of the plastid-encoded maturase MatK, RNA-binding proteins are encoded in the nucleus and imported posttranslationally into the two DNA-containing organelles. Members of two families of RNA-binding proteins have been investigated in this thesis. The two families are introduced in the following sections.

1.2.5.1 Pentatricopeptide repeat proteins (PPRs)

PPR proteins are characterized by degenerated 35 amino acid repeats arranged in arrays of tandem motifs. One PPR motif folds into two antiparallel helices and multiple PPR motifs form a superhelical extended surface (Small and Peeters 2000). This architecture places them into a superfamily of alpha-solenoid proteins, which also include other nucleic acid recognizing proteins [e.g. transcription activator–like (TAL) effector proteins and Pumilio and FBF homology (PUF) proteins]. PPR proteins, TALEs and PUF domain containing proteins recognize nucleic acids in a one repeat one nucleotide mode, with only few amino acids determining the base specificity (reviewed in Hammani et al. 2014). In PPR proteins these positions are amino acid 6 and amino acid 1 of the following repeat (position 1'), according to the nomenclature introduced by Lurin et al. (2004). Amino acid 3 seems to be involved in binding as well but whether it provides also specificity is cur-

rently unclear (reviewed in Barkan and Small 2014). A code was proposed based on frequent combinations of amino acids in position 6 and 1' and alignments of known RNA targets of PPR proteins (Barkan et al. 2012, Takenaka et al. 2013a, Yagi et al. 2013). According to the code, position 6 determines whether pyrimidine or purine bases are bound. Asparagine at this position correlates with pyrimidines and serine or threonine with purines. Position 1' helps to distinguish the two purine and pyrimidine bases (Figure 3). For less frequently co-occurring amino acids at positions 6 and 1' a nucleotide preference could not be assigned so far, due to a relatively small number of known PPR-RNA pairs. Thus there is a need for more verified PPR-RNA pairs. Amino acids interacting with RNA bases and overall structure was confirmed by crystallization of a natural and artificial PPR proteins (Coquille et al. 2014, Yin et al. 2013).

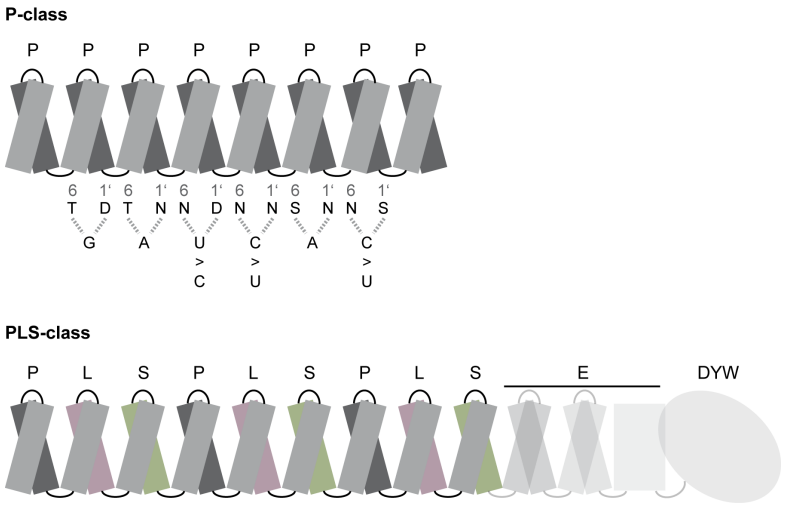


Figure 3: Two classes of PPR proteins and their mode of RNA recognition. P-class PPR proteins consist of a number of classical 35 amino acid long repeats. Each repeat is composed of two anti-parallel arranged α-helices (light and dark gray boxes). PLS-class proteins contain classical P motifs but in addition contain short (S) and long (L) motifs that are often arranged in the P-L-S order. With few exceptions PLS-class proteins carry additional C-terminal motifs. The E domain is found in almost all members of the PLS-class, and is predicted to contain four α-helices resembling two highly degenerated PPR repeats (Yagi et al. 2013). About half of the PLS proteins have an additional DYW domain that shows similarities with deaminase domains (Iyer et al. 2011, Salone et al. 2007). PPR proteins bind RNA in a one repeat one nucleotide manner. Amino acids at positions 6 and 1 in the following repeat (1') determine the specificity. Frequent combinations of amino acids 6 and 1' and their predicted binding preferences are shown below the P-class PPR model (Barkan et al. 2012, Takenaka et al. 2013a, Yagi et al. 2013). Modified from Barkan and Small (2014).

The crystal structure of PPR10 in RNA unbound and bound state showed dimeric complexes, a characteristic that was challenged by a number of studies showing that, in

solution, RNA free and bound PPR10 is monomeric under physiologically more relevant concentrations (Barkan et al. 2012, Gully et al. 2015, Li et al. 2014).

PPR proteins are found in all eukaryotes and are intimately connected to gene expression in DNA-containing organelles (reviewed in Schmitz-Linneweber and Small 2008). Members of this family in mammals are involved in mitochondrial RNA metabolism (reviewed in Rackham and Filipovska 2012). PPR proteins are dramatically expanded in land plants with more than 400 members found that, almost exclusively, are targeted to mitochondria and plastids (reviewed in Schmitz-Linneweber and Small 2008). The family in plants can by divided into two classes of similar size. Proteins belonging to the P-class are composed of canonical motifs of 35 amino acids, denoted P motifs. In contrast, members of the PLS-class contain additional motifs with a related consensus sequences. L motifs (L for long) contain 35-36 amino acids, whereas S motifs (S for short) contain 31 amino acids per repeat. They are often arranged in a P-L-S order, hence the name for this class (Lurin et al. 2004). PLS-class proteins contain, with few exceptions, additional C-terminal domains (Figure 3). About half of the proteins contain a so called E domain, with unknown function. The other half contains an E domain followed by a DYW domain (Lurin et al. 2004). The DYW domain, named after a conserved tripeptide at the C-terminus, shows similarities with cytidine deaminases and members of the PLS-class are implicated in RNA editing (Iyer et al. 2011, Salone et al. 2007). The majority of P-class PPR proteins does not contain much more than an organelle targeting sequence and an array of PPR repeats. A small group of P-class proteins contain an additional small MutS-related (SMR) domain. This domain can confer RNA and DNA endonuclease activity in different systems, but evidence for similar activity in PPR proteins is missing (reviewed in Liu et al. 2013b).

Molecular functions for a number of PPR proteins have been described. Most members of the PLS-class are implicated in recognition of *cis*-elements upstream of editing sites (see above, 1.2.4.4). P-class proteins are implicated in translation and intergenic and end processing (see above, 1.2.4.2). Exceptions from this basic rules can be found. The PLS-class protein CRR2 is involved in intercistronic processing and potentially in translation of the *ndhB* transcript (Hashimoto et al. 2003).

1.2.5.2 PPR-like proteins

Proteins with similar architecture as PPR proteins are found in chloroplasts of land plants but with lower numbers. The PPR motif is believed to have originated from the more

widespread tetratricopeptide repeat (TPR) motif (reviewed in Barkan and Small 2014). A variant of the TPR repeat, termed Half a Tetratricopeptide repeat (HAT), or previously R-TPR, is present in RNA-binding proteins in chloroplasts of *Chlamydomonas* and higher plants (Hammani et al. 2012). Members of this family (NAC2, Mbb1, and HCF107) were shown to be involved in transcript processing and RNA stability as well as translation, very similar as described for members of the P-class PPR proteins (Felder et al. 2001, Hammani et al. 2012, Schwarz et al. 2007, Vaistij et al. 2000). In *Chlamydomonas*, PPR proteins are found in small numbers but a related family is expanded in this unicellular algae. Members are predicted to fold into similar structures like PPR proteins, but the individual repeats contain 38 amino acids. The family was thus named octatricopeptide repeat (OPR) proteins. Members of this family are involved in RNA metabolism, and RNA-binding was demonstrated for individual members (reviewed in Hammani et al. 2014). Members of the mitochondrial transcription termination factors (mTERFs) are abundant in plants with about 30 members. They are composed of an array of tandem repeats and share similarity in their predicted structure with PPR proteins. They are, similarly as PPRs and OPRs, predominantly localized in plastids or mitochondria (Babiychuk et al. 2011). Even though the founding member of the family, mammalian mitochondrial termination factor 1 binds DNA, RNA-binding and a splicing defect of a chloroplast intron was shown for chloroplast mTERF4 (Hammani and Barkan 2014). In summary, PPR-like proteins have been reported to fulfill similar functions as PPR proteins in the two DNA-containing organelles of plants.

1.2.5.3 Chloroplast ribonucleoproteins (cpRNPs)

Helical repeat proteins in plastids, PPR and PPR-like proteins, bind only a few target mRNAs. In contrast, cpRNPs have been shown to bind to a variety of plastid transcripts (Kupsch et al. 2012, Nakamura et al. 1999, Nakamura et al. 2001). The cpRNP family consists of ten members in *Arabidopsis thaliana* (reviewed in Ruwe et al. 2011) and is characterized by a conserved domain structure. Two RNA recognition motifs, classical RNA-binding motifs, are preceded by a domain rich in glutamic and aspartic acid residues, thus termed acidic domain (reviewed in Nakamura et al. 2004). The cpRNP family is related to eukaryotic heterogeneous nuclear ribonucleoproteins (hnRNPs) and is not of cyanobacterial origin (Maruyama et al. 1999). In tobacco, cpRNPs have been shown to be highly abundant (Nakamura et al. 2001). Among the *Arabidopsis* cpRNPs, CP31A contains the longest acidic domain and has been shown to be additionally phosphorylated at two serine residues

in the acidic domain (Reiland et al. 2009). A spinach ortholog of CP31A binds RNA with reduced affinity after phosphorylation *in vitro* (Lisitsky and Schuster 1995). CP31A and other cpRNPs (CP29A, CP29B, and CP33B) have been shown to be phosphorylated *in vivo* (Reiland et al. 2009).

Genetic analysis of cpRNP mutants revealed a wild-type phenotype for mutants of CP31A, CP31B and CP29A under standard growth conditions (Kupsch et al. 2012, Tillich et al. 2009). T-DNA insertions in the two paralogous genes *CP31A* and *CP31B* are associated with a reduction of RNA-editing efficiency at a number of chloroplast editing sites, with effects in *cp31a* in general stronger (Tillich et al. 2009). In addition, *cp31a* mutants show reduced accumulation of several plastid mRNAs, mostly encoding subunits of the NADH dehydrogenase-like (NDH) complex. Among these, the *ndhF* mRNA is most severely affected. Transcription rates have been determined to be unchanged so that a reduction in RNA stability was assumed (Tillich et al. 2009).

Mutants of CP29A and CP31A are chlorotic under low temperatures with reduced accumulation of several plastid mRNAs and defects in RNA splicing and intercistronic processing (Kupsch et al. 2012).

1.3 Aim of this study

Processing of large polycistronic mRNAs in chloroplasts into smaller units of mono- and dicistronic mRNAs is not well understood. An endonucleolytic cleavage mechanism and lately an alternative mechanism by protein-mediated blockage of exonucleolytic activities has been proposed (Pfalz et al. 2009). The latter hypothesis predicts the accumulation of small RNAs as footprints of RNA-binding proteins (Figure 2). To investigate whether this roadblock mechanism presents the rule or the exception in chloroplasts, small RNA datasets will be investigated to identify potential footprints of RNA-binding proteins. Accompanied by precise transcript end mappings, these analysis should shed light onto the complex processing of chloroplast transcripts. In addition, such an analysis will be performed for the second DNA-containing organelle in plants the mitochondrion. How mitochondrial transcripts are stabilized in plants is relatively unclear. Analysis of small RNA datasets will give a hint whether protein-mediated protection is present in mitochondria as well.

CP31A is involved in the stabilization of a number of chloroplast transcripts. The stability of the *ndhF* mRNA is severely reduced in *cp31a* mutants. How CP31A affects the

stability of *ndhF* is a second focus of this thesis. Identification of transcript ends in wild-type and mutants could help to understand why especially the *ndhF* mRNA is so dramatically reduced in *cp31a* mutants.

RNA editing in chloroplasts and mitochondria of land plants changes several hundred genomically encoded Cs into Us on the level of RNA. Analysis of RNA editing by massive parallel sequencing of cDNAs (RNA-Seq) has not been applied to chloroplast transcriptomes yet. Quantification of RNA editing by RNA-Seq will be explored and compared to other methods used. Additional RNA-editing sites might be present in the chloroplast transcriptome and a strategy to identify these potential sites will be developed.

2 Results

2.1 Identification and analysis of small non-coding RNAs in chloroplasts and mitochondria

Previous work had suggested that stable binding of pentatricopeptide repeat (PPR) proteins to RNA could generate short RNAs in chloroplasts (Pfalz et al. 2009). In this thesis, a whole genome approach to identify small RNAs from chloroplasts and mitochondria is presented with the aim to catalog small RNAs from organelles including potential binding sites of RNA-binding proteins (RBPs).

2.1.1 Size distribution and abundance of small RNAs mapping to organelles

Analysis of small RNAs has been a focus of research over the last years in plant biology (reviewed in Voinnet 2009). Using next-generation sequencing, different classes of regulatory small RNAs have been identified, including the most prominent examples miRNAs and siRNAs (reviewed in Arikit et al. 2013). A number of small RNA datasets is available for different species and different growth conditions from public databases like the Sequence Read Archive (http://www.ncbi.nlm.nih.gov/sra/). Many of the studies investigated small RNAs using total RNA as an input. Thus the datasets include sequences derived from the nuclear genome, but also from the two DNA-containing organelles. Given the specific size of miRNAs and siRNAs (21-24nt), which are processed by distinctive machineries, many datasets have a very narrow size range. By contrast, an *Arabidopsis* small RNA dataset published by Schmitz and colleagues includes small RNAs from 15 to 50nt (Schmitz et al. 2011). The wider size distribution allows a more thorough analysis of organellar small RNAs. The sequencing results are available at the Sequence Read Archive at NCBI (SRA accession: SRA035939). The study includes eight different wild-type (WT) libraries. These were pooled before adapter sequence trimming and mapping to the *Arabidopsis* nuclear and organellar genomes using the short read aligner bowtie (Langmead et al. 2009), reporting all best alignments (4.2.19).

Using a total of 110,494,550 trimmed and quality filtered reads 33,532,813 alignments with the plastid genome were obtained. The number of reads which give rise to these alignments is 18,939,949 which represent 20.4% of all reads that could be mapped to the entire *Arabidopsis* genome (mappable reads, Table 1). Around 1.5 million reads (1.6% of mappable reads) do align with the mitochondrial genome. The largest portion of reads,

about 75 million, align with the five nuclear chromosomes resulting in ~ 430 million alignments. The discrepancy between alignments and reads is explained by the presence of two large inverted repeats in the chloroplast genome of most land plants, which encode the highly abundant rRNAs and some tRNA species. The presence of multiple copies of rRNAs and tRNAs in the five nuclear chromosomes explains the discrepancy between alignments and reads for nuclear chromosomes 1-5, as reads can map to the different copies equally well (Table 1). The total number of mapped reads translates into ~120,000 reads/kb of plastid DNA (Table 1). Fewer reads per kb of genomic sequence were found for the mitochondrial genome (4,000 reads/kb, Table 1). Even fewer reads per kb of nuclear genome sequence were obtained (650 reads/kb).

Table 1: Mapping statistics of *Arabidopsis* small RNAs using a published dataset (Schmitz et al. 2011)

chromosomes	alignments	mapped reads	reads/kb	% of mappable reads
chloroplast	33,532,813	18,939,949	~120,000	20.4%
mitochondria	1,584,688	1,478,495	~4,000	1.6%
Chr1-5	432,534,862	74,744,221	~650	80.4%

When considering the length of small RNAs it is noticeable that alignments with the mitochondrial genome are enriched in small RNAs with a length of 24nt, most likely representing siRNAs that are involved in silencing nuclear copies of mitochondrial DNA (NUMTs), (Figure 4). Alignments with the five nuclear chromosomes are enriched for sequences with 21, 23-24 and 31-33nt representing mostly miRNAs, siRNAs and tRNA fragments respectively (reviewed in Raina and Ibba 2014, reviewed in Voinnet 2009). Chloroplast alignments are enriched for 22nt reads, which can be attributed to a single RNA species mapping upstream of the *ndhB* gene encoding a subunit of the NADH dehydrogenase-like (NDH) complex. This small RNA likely represents the *in vivo* footprint of the PPR protein CRR2, since CRR2 has been shown to be essential for the intercistronic processing event overlapping this small RNA (Hashimoto et al. 2003, Pfalz et al. 2009)

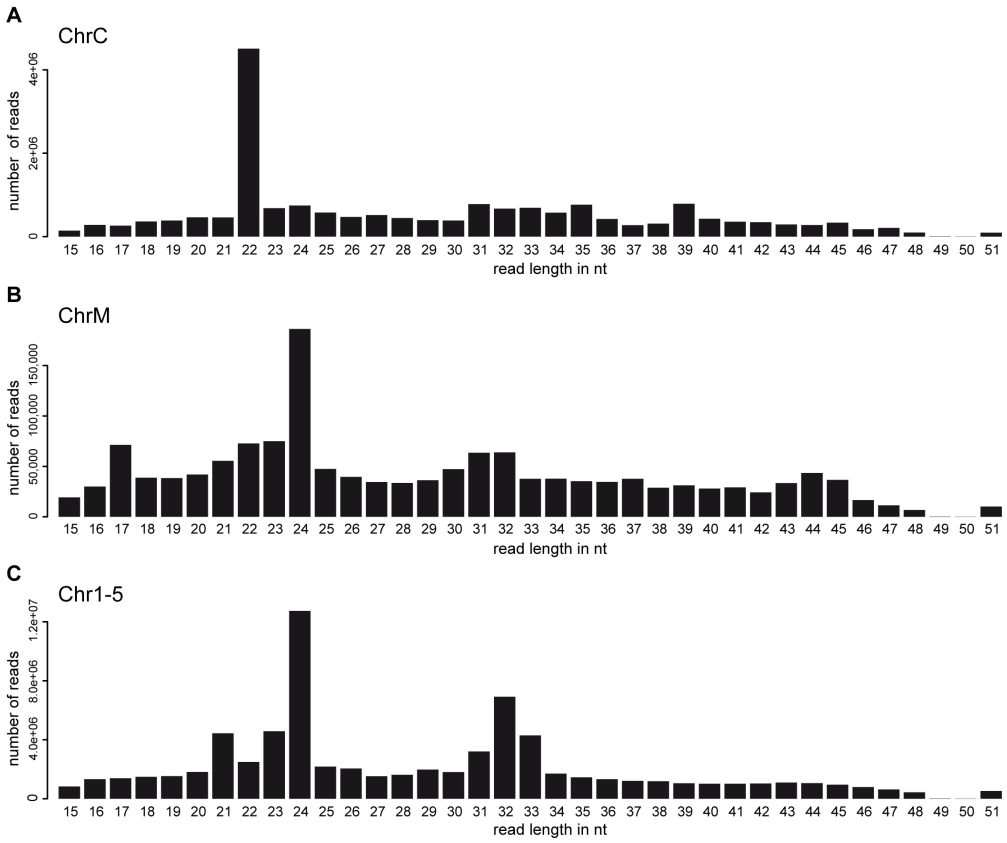


Figure 4: Size distribution of *Arabidopsis* small RNAs mapping to nuclear and organellar chromosomes. The lengths of small RNAs aligning with different chromosomes were extracted from the mappings using the SAMtools package (Li et al. 2009a). Numbers of reads obtained were plotted against the length in nucleotides. (A) Small RNAs aligning with the chloroplast genome (ChrC). (B) Small RNAs aligning with the mitochondrial genome (ChrM). (C) Small RNAs aligning with the five nuclear chromosomes (Chr1-5).

2.1.1.1 Identification of small RNAs in the chloroplast

Chloroplast transcripts differ in their abundance. Ribosomal RNAs and tRNAs are key to the organellar gene expression system and highly abundant (Legen et al. 2002). Degradation of this highly structured RNAs leads to the accumulation of degradation intermediates that include some in the investigated size range. The most abundant chloroplast mRNAs in green tissue of dicotyledons are *psbA* and *rbcL* encoding the D1 subunit of photosystem II and the large subunit of Ribulose-1,5-bisphosphate carboxylase/oxygenase (RuBisCO) respectively (Legen et al. 2002). Abundant mRNAs also produce abundant RNA degradation intermediates. To distinguish specific small RNAs from these random degradation products, an algorithm was developed together with M.Sc. Gongwei Wang,

who also implemented this algorithm. RBP protected fragments are trimmed by exonucleases and should thus have relatively sharp ends and can be separated by this characteristic from random degradation products.

The algorithm developed searches for local maxima of alignment end points in a window of 15nt with at least 40 alignments starting or ending at this position. It compares the number of alignment end points with the number found in an up- and downstream window, thus taking overall expression in the genomic region into account. If it is above a threshold described in the methods section, it is kept as a potential end of a small RNA (4.2.19). The analysis is performed for sharp 5' ends and sharp 3' ends independently, and the results are later merged. The second end of the small RNA is determined by inspecting a window of 15 to 50 bases for the most dominant 3’ or 5’ end respectively.

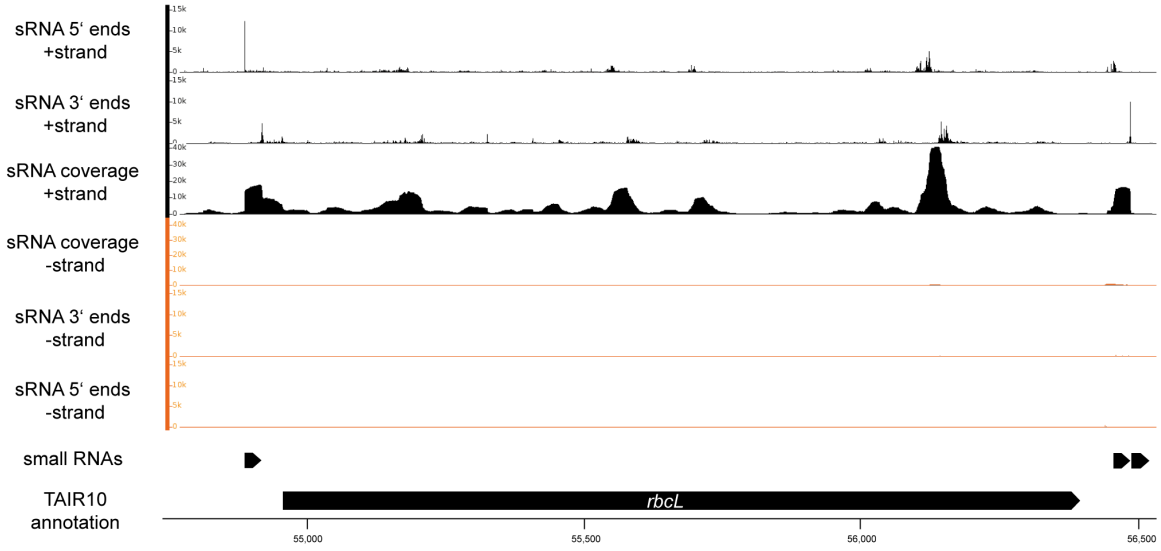


Figure 5: Identification of small RNAs in the *rbcL* region. Small RNA coverage and counts of alignment 5’ and 3’ ends is plotted against genome position and visualized using the Integrated Genome Browser (Nicol et al. 2009). Alignments with the positive strand are shown in black. Alignments with the negative strand are shown in orange. Annotations from TAIR10 (Lamesch et al. 2012) are shown on the bottom. Small RNAs identified using the developed algorithm (4.2.19) are shown above the gene annotations, with arrows indicating strandness. Three small RNAs are identified in the genomic region shown. One upstream of *rbcL* overlaps with the processed 5’ end described (Johnson et al. 2010). Two additional small RNAs are identified downstream of *rbcL*, one overlapping with a stable stem-loop and a second just downstream of that stem-loop with low abundance that is not obvious in the sRNA coverage due to the scaling of the y-axis.

Figure 5 illustrates the analysis on the example of the highly expressed plastid gene *rbcL*. The small RNA coverage is shown for both strands separately. The positive strand encoding RuBisCO shows a high coverage with small RNAs which is not present on the

negative strand. In the coding region, peaks of small RNA coverage show near normal distribution. In contrast, two peaks in the untranslated regions show sharp 5' and/or 3' ends. The peak in the 5' UTR is located at the described processing site dependent on PPR protein MRL1 (Johnson et al. 2010). The peak in the 3' UTR which shows a sharp 3' end is located at the 3' end of the *rbcL* mRNA which ends in a stable stem-loop (Zurawski et al. 1981). When plotting the starts and ends of alignments the highest count is found for these two regions even though they do not represent the highest overall coverage (Figure 5). In total, three small RNAs are identified in the region shown in Figure 5, two overlap with the processed 5' and 3' end. An additional small RNA with low coverage is found downstream of the stem-loop structure and is thus not part of the dominant mRNA encoding RuBisCO. It is found in a region were the RBP RHON1 was suggested to bind and terminate transcription of *rbcL* (Chi et al. 2014).

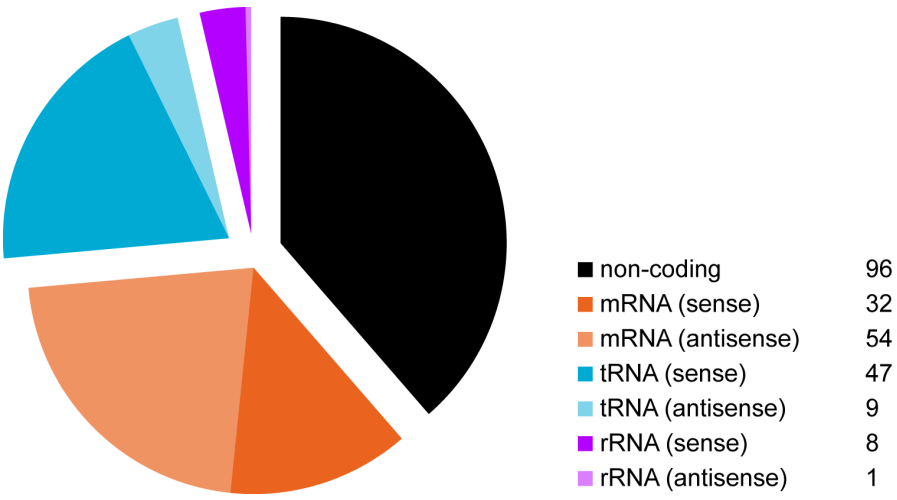


Figure 6: Distribution of small RNAs in the chloroplast genome. Overlap of small RNAs with gene annotations from TAIR10 is shown (Lamesch et al. 2012). Overlap with tRNAs (blue), rRNAs (purple) and mRNAs (orange) was identified using BEDTools (Quinlan and Hall 2010).

Using the algorithm with the parameters described in section 4.2.19, 244 chloroplast small RNAs can be identified (Figure 6, Supplementary Table 2). About one-fifth represents small RNAs overlapping tRNA annotations (Figure 6). Two types of tRNA fragments accumulate predominantly: fragments which start at RNase P processing site at the 5' end of the mature tRNA, often terminating in the anticodon stem-loop and fragments which end at the mature 3' end of the tRNA (e.g. fragments in *trnR* in Figure 7). A few small RNAs antisense to annotated tRNAs were identified. This may indicate that antisense RNAs to selected tRNAs exist that fold into stable, nucleases insensitive, structures mirroring tRNA

structure segments. Few small RNAs were identified in regions encoding ribosomal RNAs (Figure 6). They mostly overlap known processing sites having thus one sharp end and are identified when the relaxed parameters of the algorithm requiring only one sharp end are applied (4.2.19). Two-fifth of the small RNAs are found in non-coding regions. A slightly smaller fraction is sense or antisense to protein-coding genes (Figure 6). Small RNAs from these last two categories are the most likely candidates for *in vivo* footprints of RBPs. In total these two classes are represented by 180 small RNAs (Figure 6).

2.1.1.2 Plastid small RNAs cluster in intergenic regions

Many chloroplast RBPs have been described to be involved in intergenic processing of precursor transcripts (1.2.4.2). Small RNAs that represent *in vivo* footprints of these RBPs should thus accumulate in the vicinity of processing events, found predominantly in intergenic regions.

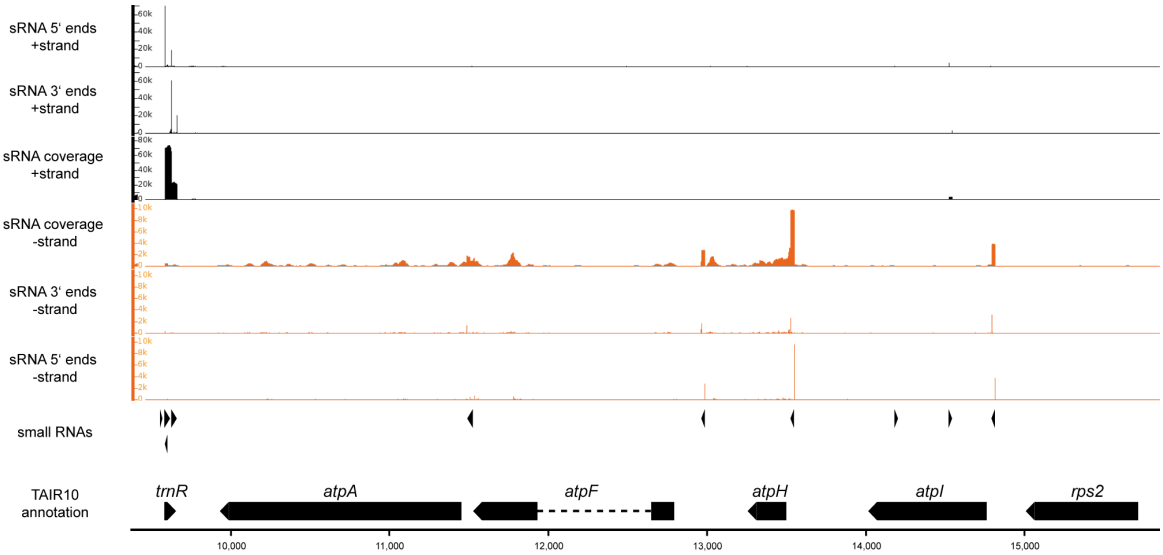


Figure 7: Small RNA accumulation in the *rps2*/*atpI*/*atpH*/*atpF*/*atpA* operon. Small RNA coverage and counts of alignment 5' and 3' ends are plotted against genome position and visualized using the Integrated Genome Browser (Nicol et al. 2009). Alignments with the positive strand are shown in black. Alignments with the negative strand are shown in orange. Annotations from TAIR10 (Lamesch et al. 2012) are shown on the bottom. Small RNAs identified are shown above the gene annotations, with arrowheads indicating orientation. Two tRNA fragments are found as well as a small RNA in the leader sequence of *trnR*-UCU. One small RNA is antisense to *trnR*, and two found antisense to *atpI*. Small RNA can be identified in every intergenic region of the operon.

Figure 7 shows the small RNA accumulation in the *rps2*/*atpI*/*atpH*/*atpF*/*atpA* operon. Inside the operon six small RNAs were identified. Every intergenic region between

two genes harbors a small RNA in sense with the up- and downstream genes. The sequence and position of the small RNA upstream of *atpH* is conserved between *Arabidopsis* and maize where it represents the *in vivo* footprint of PPR10 (Barkan et al. 2012, Pfalz et al. 2009). Two small RNAs were identified located in the *atpI* coding region but with antisense orientation. One of these small RNAs is later discussed as an *in vivo* footprint of CRR2 (2.1.4).

This pattern of small RNAs in intergenic regions is also apparent in other operons. For example in the *psbB*/*psbT*/*psbH*/*petB*/*petD* operon three small RNAs are identified in the four intergenic regions (data not shown). Only in the *psbB-psbT* intergenic region no small RNA could be identified. This finding is in line with no apparent monocistronic transcripts identified for *psbB* or *psbT* (Felder et al. 2001).

2.1.2 Transcript ends of plastid genes coincide with small RNAs

The *in vivo* footprint of PPR10 in maize was shown to overlap with processing sites (Pfalz et al. 2009). To test whether other small RNAs, which are found close to plastid genes, coincide with processing sites, transcript 5' and 3' ends were mapped for a number of genes in the proximity of small RNAs. Rapid amplifications of cDNA ends (RACE) were performed. A short RNA or DNA oligonucleotide was ligated by T4 RNA ligase 1 with total RNA. The design of the oligonucleotides allows selective ligation to either 5' or 3' ends. The sequence of the oligonucleotide serves as a binding site for a primer in a following RT-PCR. The second primer is gene-specific.

Figure 8 depicts three different examples which represent different scenarios found for small RNAs in intergenic regions. In Figure 8A the situation in the intergenic region between *rps15* and *ycf1* is illustrated. A single small RNA species is found in this region and transcript ends of the upstream and downstream cistron overlap with the small RNA. More specific, the 3' ends of the upstream cistron *rps15* overlap with the 3' ends of the small RNA and the 5' ends of *ycf1* overlap with the 5' ends of the small RNA (Figure 8A). Transcript ends thus overlap, and two precursor molecules are needed for the generation of one processed upstream and downstream transcript. This is in line with the proposed model, in which one RBP stabilizes upstream and downstream processed transcripts (Pfalz et al. 2009, Ruwe and Schmitz-Linneweber 2012, Zhelyazkova et al. 2012a).

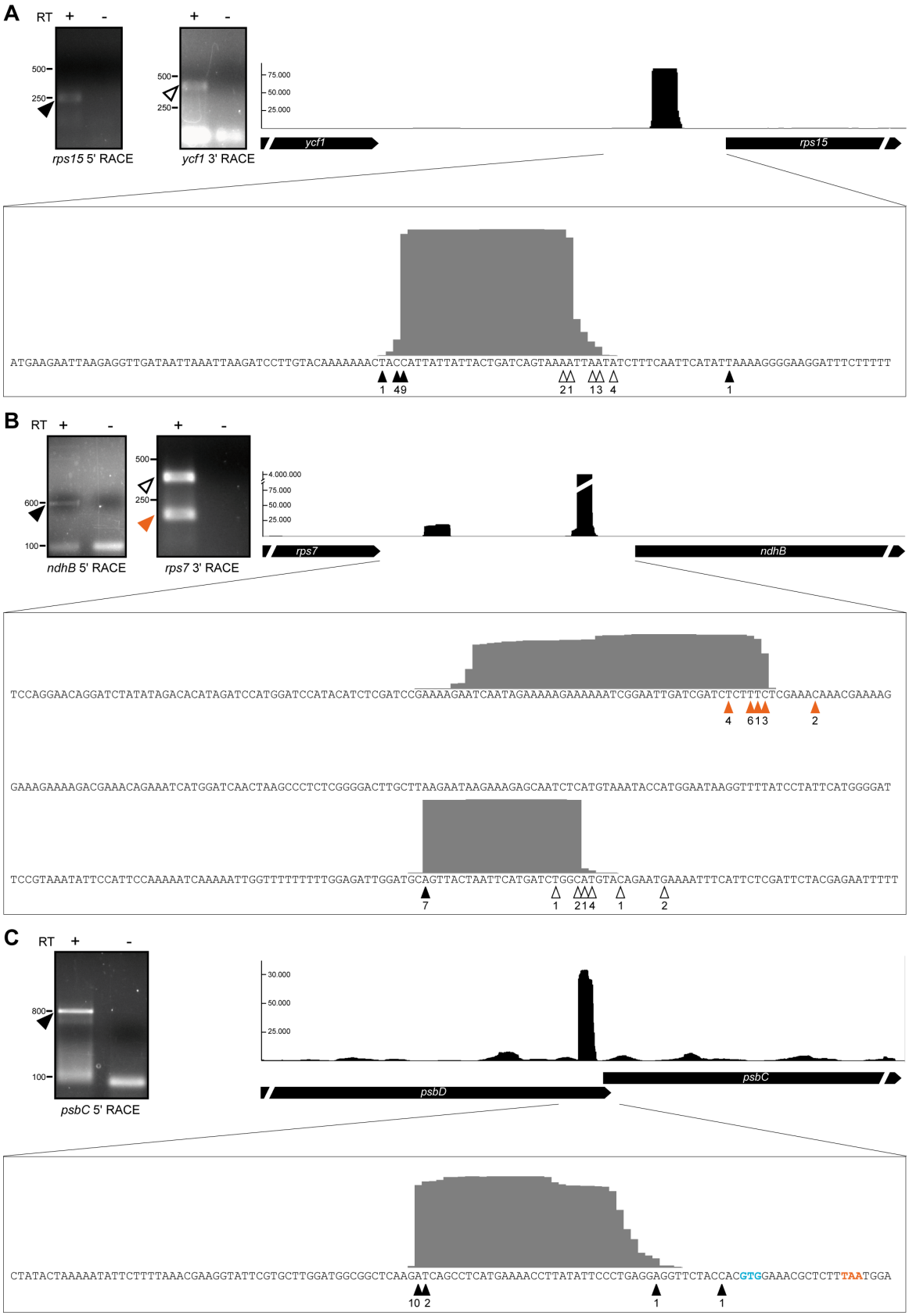


Figure 8: Transcript ends coincide with small RNAs. Rapid amplification of cDNA ends (RACE) for selected transcripts in the vicinity of small RNAs. Total RNA from WT *Arabidopsis* was ligated to RNA oligos at the 5' or 3' end. Ligated RNA was reverse transcribed and PCR amplification was

performed with an oligo-specific and a gene-specific primer. PCR products were separated on agarose gels and specific products marked by arrowheads gel-purified and cloned. Clones were sequenced and positions of 5' or 3' ends of RNAs annotated. Numbers above arrowheads correspond to numbers of clones obtained with identical 5' or 3' ends. Colors indicate the origin of clones and correspond to bands marked in the gel images. Genes are indicated by black arrows. Small RNA coverage is shown in black and in gray for the enlarged region around the small RNA identified. (A) Transcript ends determined for *rps15* and *ycf1*. (B) 3' end mapping for *rps7* and 5' end mapping for *ndhB*. Two small RNAs were identified in this intergenic region. (C) Mapping of 5' ends of *psbC*. No PCR products were obtained for 3' ends of *psbD*. The GTG triplet shown in blue corresponds to the start codon identified in tobacco (Kuroda et al. 2007). The *psbD* stop codon is marked in orange.

Figure 8B shows the situation in the *rps7-ndhB* intergenic region. Two small RNAs were identified, one overlapping the annotated start codon of *ndhB*, which represents the small RNA with the highest coverage over the whole small RNA-chloroplast alignment. The 5' end of the small RNA coincides with the site of CRR2-dependent intergenic processing (Hashimoto et al. 2003), which is here confirmed to be a dominant transcript end of *ndhB* (Figure 8B). The overlap of the small RNA with the annotated start codon likely is due to a misannotation in the chloroplast genome (NCBI: NC_000932) as the phylogenetically more conserved start codon is found 53bp downstream (Ruwe and Schmitz-Linneweber 2012). The small RNA also overlaps with 3' ends of *rps7* which is in agreement with previous data showing that beside processed *ndhB* one mRNA isoform encoding the *rps7* open reading frame is missing in *crr2* mutants (Figure 13B), (Hashimoto et al. 2003). A second small RNA is found more proximal to *rps7* which also overlaps with transcript 3' ends of *rps7*. The small RNA is one of the longest identified with the most abundant small RNA sequence 41nt in length (Figure 8B and Figure 13A). This demonstrates that in some cases, intergenic processing can lead to non-overlapping transcript ends and judging from the different length of the small RNA, that more than one RBP can be responsible for the stabilization of processed transcripts in one intergenic region.

The third example is a small RNA which is found upstream of the *psbC* gene. The *psbC* start codon is located in the upstream gene *psbD* and translational coupling of these two genes was speculated and later shown in a tobacco *in vitro* translation system. Nevertheless the *psbC* gene can also be translated from a monocistronic mRNA (Adachi et al. 2012). The 5' end of this monocistronic mRNA overlaps with the small RNA identified in *Arabidopsis* (Figure 8C). Attempts to detect 3' ends of upstream transcripts coinciding with the small RNA failed. Transcripts with such a 3' end would miss the *psbD* stop-codon. Potentially translational activity on *psbD* prevents formation of such end, i.e. ribosomes

displace the RBP from its target RNA and thus liberate the 3' ends making them susceptible for exonucleolytic degradation.

More transcript ends overlapping with small RNAs in *Arabidopsis* have been published (Ruwe and Schmitz-Linneweber 2012). The large number of reported coincidences of small RNAs with transcript ends points to a dominant role of protein-mediated protection of processed mRNAs in chloroplasts of land plants.

2.1.3 RBP dependent accumulation of small RNAs

If processing of transcripts is dependent on RBPs, and small RNAs that coincide with transcript ends represent the footprints of these proteins, these should be missing in mutants of RBPs. To verify the hypothesis, three mutants defective in specific processing events were investigated for accumulation of small RNAs at processing sites. Mutants investigated were: *hcf107*-2, *hcf152*-1 and *mrl1*-1 (Felder et al. 2001, Johnson et al. 2010, Meierhoff et al. 2003). RNA was prepared from leaf tissue using a column-based approach that recovers small RNAs. To test whether small RNAs, which are found in the vicinity of described processing sites, are disturbed in the mutants, RNA gel blots were hybridized with radio-labeled DNA oligonucleotides (Figure 9).

In mutants that are deficient for the PPR protein MRL1, a processed form of the *rbcL* mRNA is absent. As shown in Figure 5, a small RNA accumulates in the 5' UTR that overlaps with the MRL1-dependent processing site identified (Johnson et al. 2010). The small RNA has a dominant 5' end but 3' ends are dispersed over about 5nt. Accordingly, small RNAs have a length of 30-35nt. In the small RNA gel blot a signal can be observed in this size range which is not present in the *mrl1* mutant (Figure 9).

Processing upstream of *psbH* is impaired in *hcf107* mutants (Felder et al. 2001). The *HCF107* gene encodes a TPR-like protein for which homologues in maize and *Chlamydomonas* are described. Both homologues are implicated in the same processing event (Hammani et al. 2012, Vaistij et al. 2000). A small RNA with the size of 30nt was identified upstream of *psbH* (C42, Supplementary Table 2). Using an antisense probe, this small RNA can be identified in the WT and all mutants except for *hcf107*-2 (Figure 9). This is in line with findings in maize and *Chlamydomonas* where a small RNA at similar position and sequence is missing in respective mutants (Hammani et al. 2012, Loizeau et al. 2014).

The PPR protein HCF152 is implicated in intergenic processing between *psbH* and *petB*. The *hcf152* mutants show a strong decrease in cytochrome b(6)f complex levels

(Meierhoff et al. 2003). The *hcf152*-1 mutant is characterized by a T-DNA insertion in a neighboring gene of *HCF152*. The *HCF152* gene itself is not interrupted by T-DNA sequence but the expression is strongly reduced (Meierhoff et al. 2003). A small RNA identified in the intergenic region between *psbH* and *petB* is strongly reduced in the *hcf152*-1 mutant while it accumulates normally in WT and mutants including *hcf107*-2 which shows a pale phenotype as *hcf152*-1 (Figure 9). A second small RNA with slightly larger size of 30nt is also detected and is below the detection limit of the RNA gel blot in the *hcf152* mutant. Small RNAs of this size are present in small RNA sequencing datasets showing lower abundance compared to the 20nt isoform. They are extended by ten nucleotides at the 3' end.

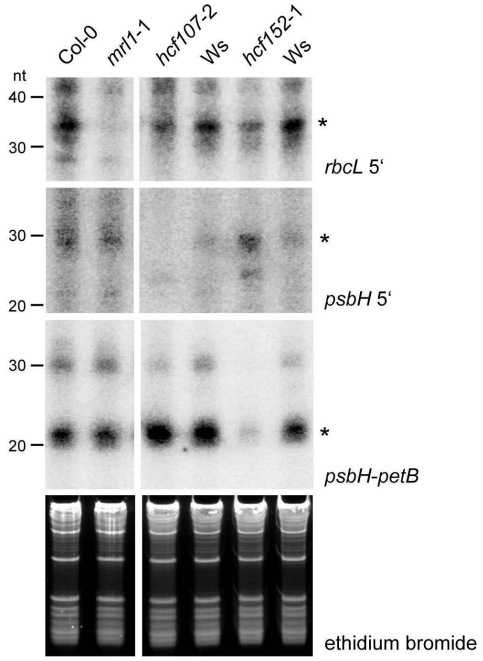


Figure 9: Small RNAs found at processing sites are missing in *mrl1*, *hcf107* and *hcf152* mutants. 3µg total RNA from the genotypes indicated were separated on denaturing polyacrylamide gels, and transferred to nylon membranes. Small RNAs were detected with ^{32}P end-labeled oligonucleotides antisense to the small RNA sequence. The approximate sizes as compared to DNA oligonucleotides are indicated in nucleotides (nt). The *mrl1* mutant and the corresponding WT in the Col-0 background were grown on soil for three weeks. Mutants with a high chlorophyll fluorescence (hcf) phenotype were grown for three weeks on MS-plates containing 3% sucrose and plants showing a pale phenotype were selected. Plants with WT phenotype were used as control (Ws). Both *hcf107*-2 and *hcf152*-1 are in the Wassilewskija (Ws) background. The ethidium bromide stain controls for equal loading. Asterisks mark bands overlapping with sizes expected from small RNA sequencing.

The absence of small RNAs in mutants of RNA-binding proteins establishes a genetic link between the presence of RNA-binding proteins and the accumulation of small RNAs.

2.1.4 Identification of RNA targets of RBPs by sequencing of small RNAs

Analysis of small RNA accumulation could serve as a quick and inexpensive way to analyze targets of RBPs, belonging to the family of helical repeat proteins. Advances in sequencing technologies allow the analysis of millions of small RNA cDNAs in few days and with relatively low costs. For a proof of principle and for potential identification of additional targets, mutants that have been described in the previous section were investigated for small RNA accumulation using sequencing of cDNAs from adapter ligated small RNAs. These include *hcf107*-2 and *mrl1*-3, mutants of two helical repeat proteins belonging to the half-a-tetratricopeptide (HAT) and PPR family respectively. Both proteins (HCF107 and MRL1) are conserved in the single-celled algae *Chlamydomonas reinhardtii* (Johnson et al. 2010, Vaistij et al. 2000). Mutants of three PPR proteins having a C-terminal SMR domain namely GUN1 (Koussevitzky et al. 2007), SVR7 (Liu et al. 2010) and SOT1 (At5g46580) were included. A *crr2* mutant (Hashimoto et al. 2003), representing a member of PLS-class PPR proteins, was included, as well as a WT of the Col-0 accession.

Using the Illumina HiSeq1500 in the rapid run mode, barcoded cDNA libraries of eleven different samples were analyzed in parallel, resulting in about 320 million reads passing filter. Four libraries are not further considered in this thesis. Two of these libraries were prepared by collaborators. One library was prepared from the *hcf152*-1 mutant that is characterized by a T-DNA insertion in the neighboring gene and is thus not a knock-out mutant. The fourth library excluded is from a T-DNA insertion in a gene encoding an uncharacterized PPR protein. The insertion is not well characterized and is located in the last exon so residual protein might be expressed. From the remaining seven, the library with the lowest number of reads was from *svr7*-3, with slightly more than 22 million reads. Reads were trimmed and mapped as described above (4.2.19). A total of 185 small RNAs were identified from the combined seven small RNA libraries using the same algorithm as described in 2.1.1.1. Of these, 148 (80%) overlap with small RNAs identified from the published small RNA dataset (Supplementary Table 2), (Schmitz et al. 2011).

For the identification of differences in small RNA accumulation between the WT and the mutants, start and end positions of alignments with the chloroplast genome were

reported including normalization per million reads mapped to the chloroplast genome. The ratio of WT and mutant samples was plotted for every position of the chloroplast genome (Figure 10). This results in four different graphs per mutant, two for each strand displaying 5' and 3' ends of the small RNA alignments (Figure 10). A high value in Figure 10 indicates that substantially more alignments start or end at this position in the WT. High ratios in the same region for 5' and 3' ends point to a small RNA missing with sharp 5' and 3' ends. Only one end with a high ratio might indicate a change in the shape of a small RNA.

Three positions in *trnR*-UCU, *trnD*-GUC and the coding region of *psbB* are found with values higher than 20 in more than one mutant. This indicates that high ratios are caused secondary or possibly even represent technical artifacts. Indeed, small RNAs in the mutants are only changed at the 5' or 3' end, which can be explained by minor technical differences in the gel elution step. They are labeled in gray in Figure 10.

The *gun1*-102 mutation (SAIL_290_D09) did not show any changes in small RNA abundance, beside the mentioned changes in tRNA regions (Figure 10C). For the *svr7*-3 mutant four changes in small RNA accumulation were observed. 3' ends of a tRNA fragment of *trnE*-UUC showed differences of one nucleotide, likely a technical artifact. Changes in the coding region of *petA* were observed, where a very low abundant small RNA is missing (Supplementary Figure 2). Two positions overlapping with 3' ends determined in this thesis, *rps7* and *ndhF*/*ycf1*as showed reduction of specific small RNA ends (Figure 10, Supplementary Figure 2). For the small RNA overlapping with a 3' end of *rps7* this leads to a change in the length distribution of small RNAs in this region (Supplementary Figure 2). Longer, 5' extended, small RNAs with a length of 40-41nt are almost absent, whereas shorter forms accumulate to about one third of the WT level. An explanation could be that two factors are responsible for the accumulation of this especially long small RNA, one being SVR7. Small RNAs that are found at the position of the major *ndhF* 3' end are in general less abundant and small RNAs with 3' extensions are absent in the *svr7*-3 mutant (Supplementary Figure 2).

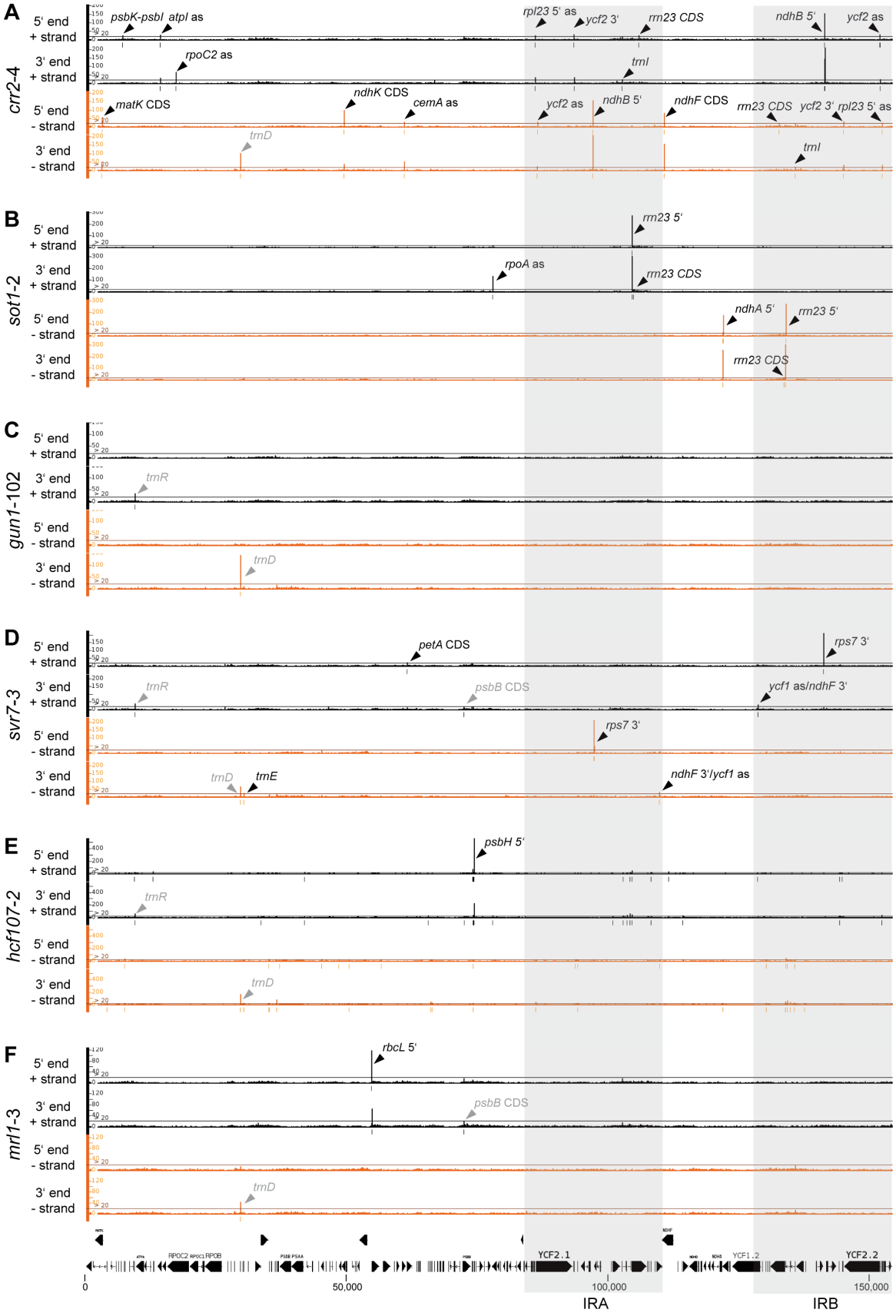


Figure 10: Identification of differential small RNA accumulation in mutants of RBPs. Small RNA libraries were prepared from three week old plants or plants in a similar developmental stage (*sot1*-2). 5' and 3' positions of small RNA alignments were extracted from small RNA mappings

using BEDTools (Quinlan and Hall 2010). Counts for every genome position were normalized to reads per million reads mapped to the chloroplast genome. Ratios of WT and mutant samples were calculated for 5' and 3' ends for positive and negative strand separately and visualized using the Integrated Genome Browser (Nicol et al. 2009). Ratios for the positive strand are shown in black and for the negative strand in orange. Ratios above 20 are indicated by small bars below the graphs. Changes specific for only one mutant are indicated by black arrowheads. When a difference in small RNA coverage was observed for more than one mutant this region is marked by a gray arrowhead. The inverted repeat regions present in the chloroplast genome are shaded in gray. (A) Differential accumulation of small RNAs in *crr2*-4, (B) *sot1*-2, (C) *gun1*-102, (D) *svr7*-3, (E) *hcf107*-2 and (F) *mrl1*-3 relative to the WT (Col-0).

In the *hcf107*-2 mutant, which has a strong phenotype, a number of changes above a ratio of 20 were observed. Many of these are likely caused secondary by the differences in RNA metabolism of the photosystem II deficient mutant, grown on sucrose-containing media compared to the WT grown on soil. Additionally it has to be noted that the *hcf107*-2 mutant is in the Wassilewskija background whereas all other mutants and the WT are of ecotype Col-0. One region with a ratio of over 400 still stands out of all other differences observed in the mutant. It is the genetically identified target of HCF107 the small RNA upstream of *psbH* (Felder et al. 2001). When compared to the *hcf152*-1 mutant grown under similar conditions and of same ecotype only three regions showed differential coverage above a threshold of 20: *psbH* 5', *rrn23* 5' and *trnD* (Data not shown).

In the *mrl1*-3 mutant small RNAs in the 5' UTR of *rbcL* are absent. They overlap with the processed 5' end described previously (Johnson et al. 2010). No additional small RNAs were found to be changed above the threshold of 20 specifically in the *mrl1*-3 mutant.

In conclusion, genetically identified targets of HCF107 and MRL1 could be confirmed by sequencing of small RNAs in mutants, demonstrating the potential of this novel technique. Analyses on small RNA accumulations in *sot1*-2 and *crr2*-3 mutants lead to additional experiments that are described in the following two sections.

2.1.4.1 PPR-SMR protein SOT1 stabilizes three small RNAs

The point mutant *sot1*-1 was isolated in a suppressor screen for a variegated leaf phenotype in the *thf1* mutation in Jirong Huang lab at the Shanghai Institute for Biological Sciences. The mutation was mapped to the gene At5g46580 which in turn was named *SOT1* for *suppressor of thf1*. A T-DNA insertion in the *SOT1* gene was isolated in the lab of Prof. Ian Small. The *sot1*-1 and *sot1*-2 have been found to be defective in plastid ribosomal RNA

maturation, especially at the 23S ribosomal RNA, where a precursor is not accumulating in *sot1* mutants (Dr. Kate Howell, personal communication).

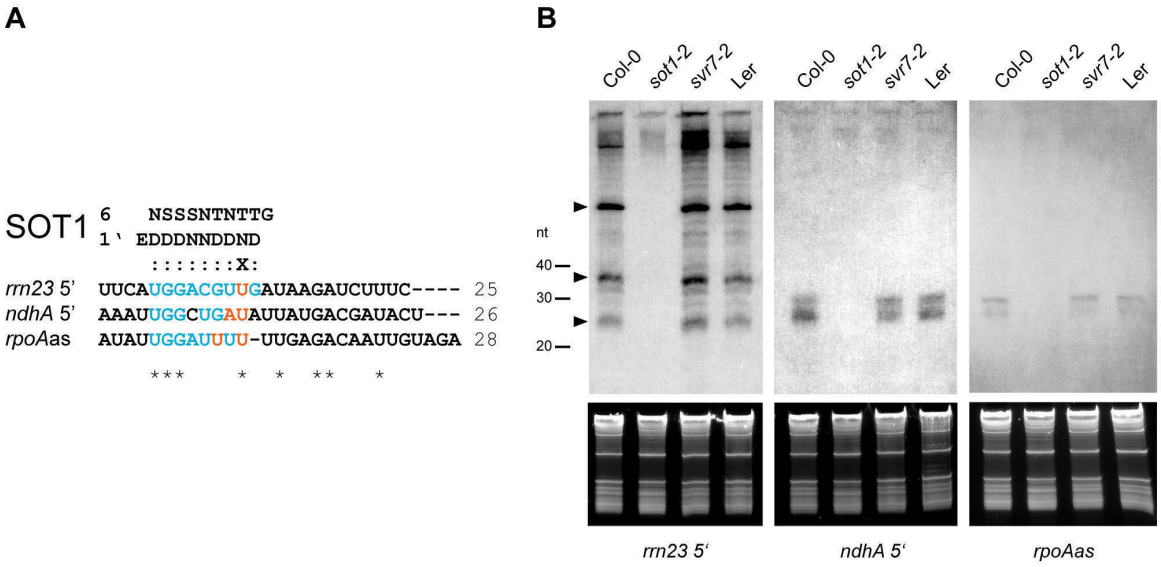


Figure 11: Three small RNAs are missing in *sot1*-2 mutants. (A) Small RNAs overlapping with regions of strong change in small RNA coverage in *sot1*-2 mutants were aligned with Clustal W2 (Larkin et al. 2007). Asterisks below the alignment indicate identical nucleotides in all three small RNAs. The amino acids at positions 6 and 1' of the PPR repeats from SOT1 are shown and aligned manually with the small RNA sequence upstream of *rrn23*. Nucleotides in blue are positively correlated with the 6/1' combinations of amino acids in individual PPR repeats, orange nucleotides negatively correlated (Barkan et al. 2012). (B) Small RNA gel blot analysis of the three small RNAs identified. 5µg RNA from each genotype was separated on denaturing polyacrylamide gels and transferred to nylon membranes. The *sot1*-2 line is in the Columbia background (Col-0) and *svr7*-2 in Landsberg erecta (Ler) background. Small RNAs were detected using end labeled oligonucleotides antisense to the small RNAs. An ethidium bromide staining is shown to control equal loading. Arrowheads mark three hybridization signals obtained with the *rrn23* 5' probe that are described in the text.

Accumulation of small RNAs differs at three genomic regions in *sot1*-2 mutants. They are located upstream of *rrn23* and *ndhA*, one is antisense to *rpoA* (Figure 10B and Figure 11). An alignment of three small RNAs found at these positions is shown in Figure 11A. The three small RNAs slightly differ in length. Three nucleotides (UGG) starting at position five of the alignment are found in all three small RNAs. These are positively correlated with amino acid combinations 6 and 1', amino acids that interact with the RNA bases, in PPR repeats 1-3 of SOT1 (see 1.2.5.1 for an introduction into the "PPR code"), (reviewed in Barkan and Small 2014). RNA gel blot analyses support the findings from small RNA sequencing. All three small RNAs were not detectable in the *sot1*-2 mutant, whereas in WT and a *svr7* mutant the small RNAs could be identified (Figure 11B).

Using a probe to detect the small RNA upstream of *rrn23*, additional SOT1 dependent bands were obtained. A band at approximately 35nt likely corresponds to an isoform of the small RNA with a 3' extension of about 10nt which is also present in small RNA sequencing datasets. The abundance seems to be reversed in the RNA gel blot compared to small RNA sequencing. An additional hybridization signal was obtained at ~75nt (all three bands marked with triangles in Figure 11). The small RNA upstream of *rrn23* does overlap with the 5' end of a precursor of the 23S ribosomal RNA (Bollenbach et al. 2005). 5' RACE analysis showed that the processed 5' end is absent in *sot1*-2 mutants, suggesting that SOT1 stabilizes this precursor (Dr. Kate Howell, personal communication and Supplementary Figure 1).

The small RNA upstream of *ndhA* overlaps with a primary transcript end as determined by 5' RACE. This transcript end, which was only amplifiable if RNA was treated with tobacco acid pyrophosphatase to covert the primary triphosphate end to a monophosphate end, is also present in *sot1*-2 mutants (Supplementary Figure 1).

2.1.4.2 Eleven small RNAs are missing in mutants of the DYW-PPR CRR2

In *crr2*-4 mutants 13 regions showed a strong reduction in small RNA coverage (Figure 10A). Two are located in the non-coding RNAs *rrn23* and *trnI*. However, no small RNAs were identified overlapping these positions. Ten other regions overlap with small RNAs. Strikingly, eight out of ten small RNAs are 24nt in length (Figure 12A). One remaining difference in the intergenic region between *psbK* and *psbI* was manually curated and does overlap with a potential small RNA that by its low abundance did not get detected by the algorithm. Strikingly it is 24nt in length. An alignment based on 5' ends of all small RNAs is shown in Figure 12A. Positions 4-18 in the alignment show sequence similarity (Figure 12B). If these small RNAs represented *in vivo* footprints of CRR2 these positions likely would represent the region of RNA-protein interaction. At positions 8 and 9 all small RNAs have an adenosine. This aligns well with PPR motifs six and seven of CRR2 (Figure 12B). Adenosine is the preferred base for P- and S-type PPR motifs with a T/N and S/N combination at positions 6 and 1' in PPR repeats (Barkan et al. 2012). Interestingly at position 15 of the alignment all sequences share a guanosine nucleotide which is outside of this alignment of small RNAs with PPR repeats (Figure 12B). Furthermore, positions 16-18 are similar between the eleven small RNAs suggesting that they provide specificity interacting with CRR2.

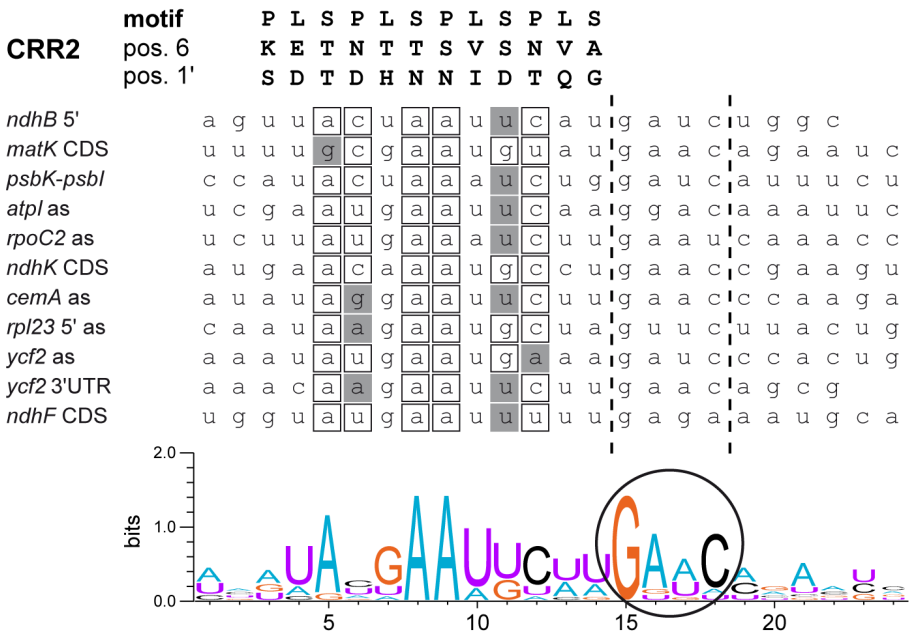


Figure 12: CRR2-dependent small RNAs are conserved in length and sequence. Eleven small RNAs that showed strong reduction in *crr2*-4 as determined by small RNA sequencing (Figure 10) are shown. A sequence logo of an alignment of all small RNAs is shown. The logo was generated using weblogo3 (Crooks et al. 2004). The amino acids at positions 6 and 1' of the PPR repeats from CRR2 are shown and aligned manually with the small RNA sequences. Boxed residues indicate a match, gray shaded residues a mismatch, with regard to the PPR code (Barkan et al. 2012). Four bases which show strong similarity between all small RNAs but are outside the alignment with the PPR repeats are circled.

To complement the results obtained from small RNA sequencing, RNA gel blot analysis was performed on two independent T-DNA insertion lines interrupting the *CRR2* gene (Table 6). The insertions were confirmed by PCR analysis (data not shown). Total RNA from homozygous mutants was separated in polyacrylamide gels and detected with end-labeled oligonucleotides antisense to the respective small RNAs. The small RNA analysis is shown in Figure 13A. Three small RNAs which had the highest number of reads in small RNA sequencing were analyzed. A second small RNA in the *ndhB-rps7* intergenic region was included as a control (*rps7* 3') and indeed is not affected or might even be increased in abundance in *crr2* mutants (Figure 13A). Three small RNAs namely *ndhB* 5', *matK* CDS and *ycf2* 3' are absent or at least decreased below the detection limit in *crr2* mutants indicating they represent footprints of CRR2.

The absence of the *ndhB* 5' small RNA in *crr2* mutants goes in hand with defects in stabilization of processed *ndhB* and *rps7* transcripts (Hashimoto et al. 2003), (Figure 13B). To test whether similarly the absence of other small RNAs in *crr2* mutants is linked to processing/stabilization defects, RNA gel blot analyses were performed to detect longer

RNA species after formaldehyde agarose gel electrophoresis. Strand-specific RNA probes were used that span the small RNAs *ndhB* 5', *matK* CDS, *ycf2* 3' and *ycf2*as. Each probe is of approximately 350nt in length, allowing hybridization with regions upstream and downstream of the small RNA. CRR2 dependent bands should thus be missing in the two *crr2* mutants. The described defect in *ndhB* and *rps7* processing is readily visible in Figure 13B. Hybridization with other probes complementary to regions where CRR2 dependent small RNAs were detected, did not result in *crr2* dependent changes in the banding pattern (Figure 13B).

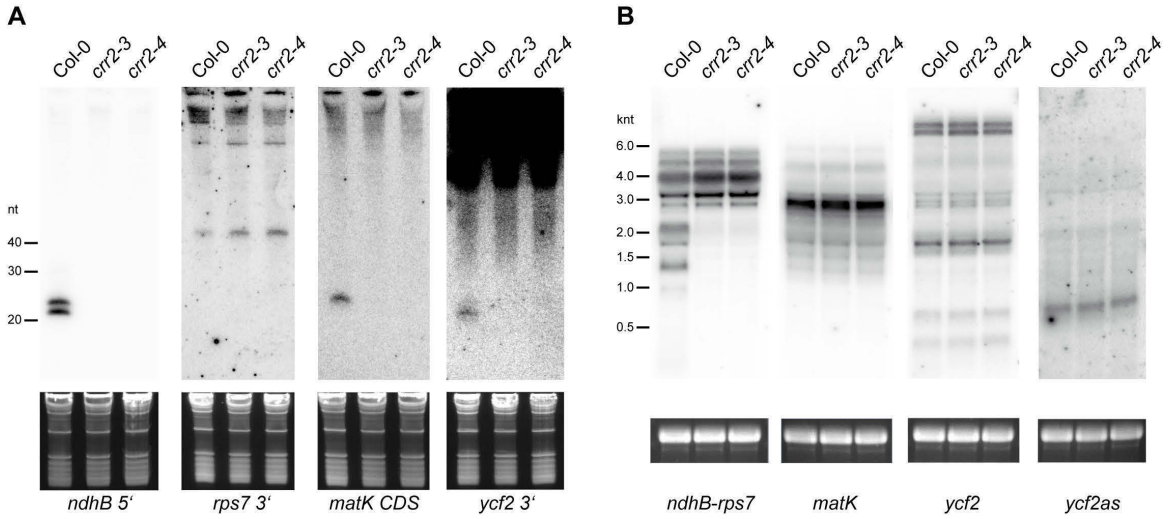


Figure 13: Analysis of RNA accumulation in *crr2* mutants. (A) Analysis of small RNA accumulation in *crr2*-3 (SALK_030786) and *crr2*-4 (SALK_046131) mutants and the WT (Col-0). 5µg RNA was separated on denaturing urea-polyacrylamide gels and blotted to nylon membranes. EDC-cross-linking was followed by hybridization with 5' radio-labeled DNA oligonucleotides complementary to small RNAs identified missing in *crr2* mutants. The small RNA *rps7* 3' is located in the *rps7-ndhB* intergenic region as is *ndhB* 5' and serves as a control. (B) Analysis of RNA accumulation by RNA gel blot analysis in formaldehyde agarose gels. 5µg RNA was loaded and transferred to nylon membranes and UV-cross-linked. Strand-specific RNA probes used extend the region of the small RNAs that are missing in *crr2* mutants by about 150nt in both directions. Ethidium bromide staining of gels is shown as a loading control.

It can be concluded that in *crr2* mutants a number of small RNAs, including a genetically identified target upstream of *ndhB*, is missing that show sequence and length similarity. These likely represent footprints of PPR protein CRR2. The conservation of small RNA sequence between the different CRR2 targets extends beyond the region predicted to be recognized by PPR repeats in 3' direction. The absence of CRR2 does not result in specific processing defects at the new binding sites identified.

2.1.5 PPR10 is bound to the small RNA upstream of *atpH*

Even though the accumulation of small RNAs that represent footprints is dependent on RBPs, it is unclear whether small RNAs exist only in a complex or alternatively also in a protein-unbound state.

To test association of a small RNA with its cognate RBP an RNase protection assay was combined with immunoprecipitation of the PPR protein PPR10 from maize stroma fractions. In the RNase protection assay a radiolabeled antisense RNA probe was hybridized with the RNA sample and single-stranded RNA i.e. non-hybridized regions of the probe were digested with single-strand specific RNases. RNAs of different sizes accordingly give rise to protected fragments of different length. Figure 14B shows the probe used to detect the small RNA, unprocessed precursor transcripts and transcripts which are processed in the *atpI-atpH* intergenic region. Hybrids consisting of the different RNA species and the labeled probe differ in length. After RNase digestion the protected fragments can be separated by size in denaturing polyacrylamide gels.

A benefit of this technique is the simultaneous detection of the different RNA species in one assay, which allows comparative estimation of small RNA and mRNA abundance. Figure 14A shows the results from the RNase protection assay. Two samples using Yeast RNA serve as controls. The sample without RNase digestion controls probe integrity during the assay. The second sample which is incubated with RNases will only give raise to signals which result from self-protection, e.g. stable structures in the probe itself. Using total maize RNA isolated from the first leaf, four strong bands are detectable corresponding to the sizes expected for precursors, two processed transcripts and small RNAs (Figure 14). The protected fragment presumably corresponding to the small RNA shows similar abundance as polycistronic precursors and processed *atpI* transcripts. It has to be taken into account that different protected fragments contain different numbers of radiolabeled nucleotides, in this case UTP, leading to an underestimation of processed *atpI* and the small RNA in the radiographs (Figure 14B). The band corresponding to processed *atpH* mRNAs represents the strongest signal which is in line with previous reports showing processed *atpH* transcripts are more abundant than processed *atpI* transcripts (Pfalz et al. 2009). In conclusion, the small RNA representing the footprint of PPR10 accumulates to similar levels as *atpH* and *atpI* mRNAs.

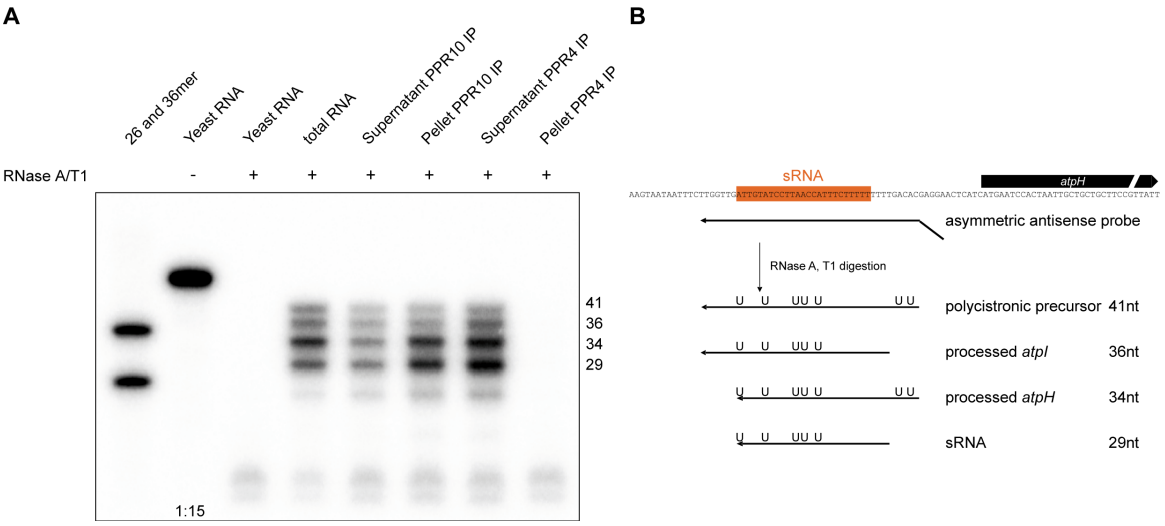


Figure 14: RNase protection experiments identify RNAs species that co-precipitate with PPR10. (A) RNase protection assay using total RNA from the first leaf of 10 day old maize seedlings and RNAs co-precipitated with PPR10 from maize stroma. 1µg of total RNA and RNA isolated from supernatants was used. For pellet fractions same partial volumes were used. RNAs were hybridized at 42°C with a ^{32}P-labeled antisense RNA and non-hybridized regions of the probe were digested with a mixture of RNase A and RNase T1. Two end-labeled RNA oligos are included as size markers. Hybridization with yeast RNA controls for probe integrity during the experiment (-RNase, 1:15 dilution) and self-protection of the probe (+RNase). Immunoprecipitation using specific antibodies for PPR10 (Pfalz et al. 2009) and PPR4 (Schmitz-Linneweber et al. 2006) was performed, the latter representing a non-specific control precipitating an unrelated PPR protein. (B) Schematic representation of the RNase protection assay. The location of the small RNA identified in maize upstream of *atpH* is shown in orange (Pfalz et al. 2009). Parts of the probe that are encoded within the chloroplast genome are aligned with the sequence. A short artificial sequence at the 5' end of the probe is not aligned. Protected fragments predicted originating from different RNA species are shown with the position of radiolabeled uridines indicated.

All four transcript forms detected in total RNA are present in pellet fractions after PPR10 immunoprecipitation. In addition, a slightly smaller band around 24nt is also found, but its origin is unclear at present. The small RNA and the processed *atpH* seem to preferentially co-precipitate as judged from two independent experiments. The majority of small RNA and *atpH* mRNA species are precipitated, compared to about half of the polycistronic precursors and monocistronic *atpI* mRNA (signals in Figure 14 can directly be compared as dilution of pellet and supernatant fractions are identical). It is unclear whether this is due to preferential binding of PPR10 to these RNA species or due to more efficient precipitation (small RNA and monocistronic *atpH* mRNA are smaller than the precursor and the processed *atpI* mRNA). As a control, an immunoprecipitation of an *atpH-atpI* unrelated PPR protein, PPR4, was included. PRR4 was shown to bind to the trans-spliced intron of *rps12* and facilitates *rps12* splicing (Schmitz-Linneweber et al. 2006). All

RNA forms remained in the supernatant, showing that the co-precipitation with PPR10 is specific.

The analysis showed that the majority of small RNAs representing the *in vivo* footprint of PPR10 are bound by PPR10. Whether protein-unbound small RNAs exist *in vivo* cannot be concluded from the data as the immunological detection of PPR10 in the pellet and supernatant fractions failed. The efficiency of immunoprecipitation thus cannot be quantified. In general the small RNA accumulates to substantial amounts in plants, comparable with the abundance of processed mRNAs judged from accumulations in total RNA.

2.1.6 Mitochondrial small RNAs

A huge number of RBPs is predicted or was shown to be imported in mitochondria of land plants (reviewed in Hammani and Giege 2014). PPR proteins represent the RBP family with the highest number of members predicted to be localized to mitochondria (Colcombet et al. 2013). Similar to plastids, mitochondrial RNAs undergo a number of RNA processing steps including end processing (reviewed in Hammani and Giege 2014). An analysis of small RNAs from mitochondria can be expected to be useful for the prediction of binding sites of RBPs similar to what was demonstrated above for plastids.

2.1.6.1 Identification of small RNAs in mitochondria

For the identification of small RNAs from mitochondria the algorithm described in 2.1.1.1 was slightly modified, as pilot analysis had shown that mitochondrial small RNAs have less well defined ends. In detail, more alignments starting in the sequence of a potential small RNA were allowed (4.2.19). Using these settings a total number of 315 small RNAs were identified (Supplementary Table 3). Of these, 119 had a length of 24nt potentially representing abundant siRNAs originating from NUMTs. This bias was not observed for plastid small RNAs (Supplementary Figure 3).

2.1.6.2 Small RNAs coincide with termini of mitochondrial transcripts

In *Arabidopsis*, mitochondrial transcript ends of protein-coding RNAs have been mapped systematically (Forner et al. 2007). To investigate whether small RNAs overlap with these transcript ends similar as in plastids (2.1.2), positions of small RNAs were analyzed with regard to the processing sites described by Forner et al. (2007). In Table 2, all small RNAs that overlap with the described transcript ends are listed. The overlap of small RNAs with processed 3' ends is more prominent than with described 5' ends. About 70% of 3' ends mapped in *Arabidopsis* overlap with a small RNA identified in this thesis. In contrast, only 14% of 5' ends determined overlap with the accumulation of a small RNA (Table 2). Strikingly none of the small RNAs overlapping transcript ends displays a length of 24nt, indicating they do not represent nuclear-encoded siRNAs (Table 2). Two 5' processing sites overlapping with small RNAs have been postulated to result from RNase Z cleavage at upstream located tRNA genes or so called t-elements (Forner et al. 2007), (Table 2). A small RNA overlaps the major 5' transcript end of *atp8* for which a conserved promotor element was found in the upstream region (Kuhn et al. 2005), (Table 2).

A number of PPR proteins have been implicated in end processing of mitochondrial mRNAs. RPF1-7 identified in *Arabidopsis* and MPPR6 identified in maize are involved in 5' processing of different mitochondrial transcripts (Binder et al. 2013, Hauler et al. 2013, Holzle et al. 2011, Jonietz et al. 2011, Jonietz et al. 2010, Manavski et al. 2012, Stoll et al. 2014, Stoll et al. 2015). Only processing sites in *atp9* and *nad6* which are decreased in *rpf5* mutants show an overlap with a small RNA (Hauler et al. 2013), (Table 2). Interestingly, the binding site of RPF5 was predicted to be located ~40-50nt upstream of the processing sites affected in *atp9*, *nad6* and *rrn26*. Thus if this prediction is correct, the processing cannot be explained by blockage of a 5'→3' exonuclease activity by RPF5 and the small RNA does not represent the footprint of RPF5. The predicted binding site of RPF5 is present in a small RNA upstream of the *rrn26* gene (M17 in Supplementary Table 3). The accumulation of a precursor of *rrn26* is increased in *rpf5* mutants, whereas the mature *rrn26* accumulates to lower levels.

The PPR protein MTSF1 was shown to be indispensable for stable accumulation of the mature *nad4* mRNA (Haili et al. 2013). A small RNA can be identified at the mature 3' end of *nad4* (Table 2). The small RNA is absent in *mtsf1* mutants recapitulating the situation for mutants of plastid localized PPR proteins (Haili et al. 2013).

In contrast to plastids, mitochondria seem to use protein-mediated stabilization predominantly at 3' ends of mRNA.

Table 2: Small RNAs identified in mitochondria overlapping transcript ends.

Gene	Flanking sequences of transcript ends identified by (Forner et al. 2007). Transcript ends are underlined. Sequences of small RNAs are in bold and blue. The major mRNA end is indicated by a large letter.	Comment
3' ends of mRNAs overlapping with small RNAs (out of 27 described 3' ends)		
nad5	CCAGGCGCC**CATTCCCAGTTCTTTCTCTTCTCTCTTTTT**T_TAGTTTAGTG	
nad9	TAGGTCCA**CCAGTCCAGGGGACAAATCAATAGGAAATGCTA**T_AGGAAATG	
ccmB	AATGTTGGGCCGGGT**ATGTAAGCCATGTATCTAGG**A_GGAATTAGAAAGAA	
ccmF$_C$	AATAGGAAA**GCTTTCAATCAATAGAAATCGTATTCGT**GAATAAATCCCCT	
cob	AT**TCTGACACCAATCATTTACATATTACACCAAGAATTGACAAGCAG**A_TA	
nad6	GATTTTAGGAGG**ACTATAATGAGGAGGACGACTGACC**C_ACTCACGATCTA	
rps4	GGCCGAGAAT**CTTATGTCAAAAGGACCAAGGACGATC**T_TTTCGGAAAGGA	
atp8	GCCTTCGCGGTTCG**ACTTTCTTTTCAGGCTTGACT**C_ATTCGCTAGCTTCT	
nad7	CTA**GTTGCTCGATCAGGACCTTAGCTTTATTGCGAGCCCAGAAGTC**T_CTC	
nad1/atp9	CGAAAATGCCCGTTAA**TCAAGCAAGTTGGGGA**A_CAAAATCTTCCTTGTTA	
mttB	AAGAGTAGCCCCC**CCCCTAGAACCTGGCAAAGTAA**C_TATCAATGAATTCC	
nad4	TTGAGAGGAATCAGCA**AAGAAAAGAAAAACGGG**T_CAACATCTTAATGTGT	3'end dependent on MTSF1[a]
atp4	ATGTTCA**TGCTCTCAGAAGAGCGGATCCAATACCAAGACTAC**T_TCTTTCT	
ccmC	A**ACGGAAGAAATTGAAGCTCGAGAAGGAATACCAAAACCTAGTTCACT**C_G	
ccmF$_{N2}$	TTTGA**TCAGTAGATTATTTAGAACTTCGGAAGATGGTCAAGGTA**C_GAAGT	
rps7	GGGAGCTGA**TCTGATAAATGCACTTCAAAGGGAGGGAAGG**C_TAGGAATCT	
nad2	TCTTTAA**GTTCGATCATTGACAAGGTTCAAAGAAAGGGTAG**G_CCGTCGGT	
cox1	AA**GAAGAAAAGGTCGCCGACTGCTACTAAGAACCTAACAGAACTTT**T_AGA	
5' ends of mRNAs overlapping with small RNAs (out of 42 described 5' ends)		
atp9	CGCAA_AGAATGCATTCCAAGTGAGATGTCCAAGATCAAAGGAACGAGGGT	processing enhanced by RPF5[b]
atp8	TATCAATCTCATAAGA**G**AAGAAATCTCTATGCCCCCTTTTTCTTGGTTTT	conserved promoter element[c]
nad6	GAAAA**G**AATGCATTAAATGGATGCATTGAGATTCCGTAAGTAACTCAGTG	processing enhanced by RPF5[b]
cox2	G_AA**GAAGAATCTTACGCCCCAAATTCCCATCTCTTTTTTCTTGGTTGGAC**	
ccmF$_C$	CTTCCGCT**C**CTGGTGTTCGAACTAGTCATTAATGGTCGGCTTCATTGGTA	end created by RNase Z
rps4	GGACGCAA**T**GTGGCTGCTTAAAAAACTGATTCAACGAGATATAGATTTGT	t-element, RNase Z?[d]

[a] The small RNA and the processed 3' end are absent in *mtsf1* mutants (Haili et al. 2013)

[b] 5' processing of *atp9*, *nad6* and a precursor *rrn26* was shown to be decreased in *rpf5* mutants (Hauler et al. 2013)

[c] A conserved promotor element is present upstream of the 5' end identified (Kuhn et al. 2005)

[d] A structure upstream of the processing site forms a structure similar to a tRNA and is potentially recognized by RNase Z (Forner et al. 2007)

2.1.6.3 Mitochondrial small RNAs have less defined 5' ends

During the analysis of small RNAs that overlap with transcript ends it was striking that small mitochondrial RNAs had less defined 5' ends. This is exemplified in Figure 15 where two mitochondrial and two chloroplast small RNAs are compared. The two chloroplast small RNAs in Figure 15 were selected, as they are located at the end of transcriptional units, thus resembling the situation in mitochondria where polycistronic

transcripts are rare. Binding of an RBP at these sequences is likely not required for stabilization of downstream sequences. As an indicator of 5' end sharpness the coverage decrease at the 5' end was measured as number of nucleotides required for a drop in coverage from 80 to 10% of maximal coverage in the region of the small RNA (Figure 15). When the sharpness of all small RNAs overlapping processed 3' ends (Table 2) was calculated, on average 14±5bp (SD) were required for this drop in coverage. This drop was significantly sharper for five chloroplast small RNAs at termini of transcription units, namely small RNAs downstream of *ndhJ*, *ycf3*, *ndhF*, *rps18* and *ycf2* (5±4SD; p=0.007 in a two tailed unpaired students t-test).

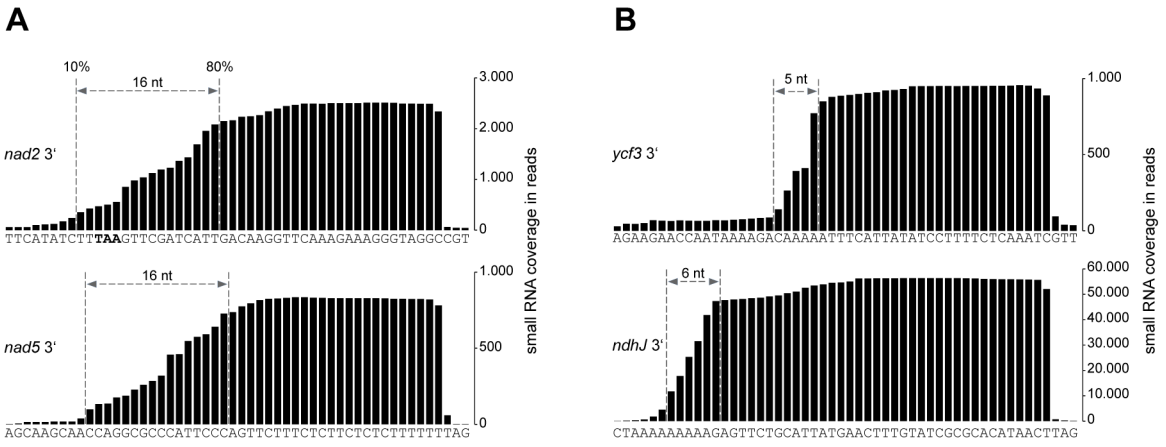


Figure 15: 5' ends of small RNAs found at 3' ends of transcripts are less defined in mitochondria. Coverage plots of small RNAs found downstream of transcription units. The coverage is shown in number of reads. The number of nucleotides required for a drop in coverage from 80 to 10% is indicated between dashed lines. (A) Two mitochondrial small RNAs that overlap with transcript ends of *nad2* and *nad5* respectively. (B) Two small RNAs downstream of the *ycf3* gene and of the *ndhC*/*ndhK*/*ndhJ* operon are shown.

2.2 CP31A stabilizes the *ndhF* mRNA by interaction with its 3' UTR

Are only PPR and PPR-like proteins involved in the generation of small organellar RNAs? Other classes of RBPs might similarly leave footprints or stabilize the PPR-RNA complexes. A potential candidate for such a non PPR protein is CP31A. CP31A is a member of a small family of RNA-binding proteins named cpRNPs. Members of this family consist of two RNA-recognition motifs that are able to bind RNA and in addition harbor an acidic domain in the N-terminus (reviewed in Ruwe et al. 2011). In *cp31a* mutants several RNA-editing sites show reduced RNA editing and especially mRNAs encoding subunits of the NADH dehydrogenase-like (NDH) complex are reduced in abundance (Tillich et al. 2009).

CP31A and a close relative CP29A are required for cold tolerance of *Arabidopsis*. Mutants of both RBPs exhibit reduced levels of several chloroplast mRNAs in the cold (Kupsch et al. 2012). Mechanistic details on how CP31A influences RNA stability are not known. Results from an analysis of the most strongly reduced mRNA *ndhF* under normal growth conditions are presented in the following sections.

2.2.1 The dominant 3' end of *ndhF* mRNA is not detectable in *cp31a* mutants

In *cp31a* mutants, the *ndhF* mRNA is reduced below the detection limit in RNA gel blot experiments (Tillich et al. 2009). The transcription rates in *cp31a* mutants are comparable to the WT (Tillich et al. 2009). Therefore, RNA stability is likely reduced in *cp31a* mutants. As demonstrated in previous sections, differences in the stability of specific transcripts is often accompanied by processing defects. The dominant 5' end of *ndhF* is primary and is located 320bp upstream of the *NdhF* start codon (Favory et al. 2005). 3' ends for *ndhF* have not been identified so far. In this thesis, 3' ends of *ndhF* transcripts were identified by rapid amplification of cDNA ends (RACE), using a linker ligation strategy (Figure 16). Total RNA was ligated to a small phosphorylated RNA oligonucleotide using the T4 RNA Ligase I. RNAs were reverse transcribed using a primer complementary to the small RNA adapter. RT-PCRs were performed using a gene-specific primer and a primer matching the adapter sequence. Figure 16 shows the PCR results for the *ndhF* 3' RACE in the WT and in a *cp31a* mutant. The dominant PCR product found in the WT is missing in *cp31a*. Other PCR products present in the WT are readily detectable. The WT specific PCR product was gel-purified, cloned and individual clones sequenced to analyze the distribution of 3' ends. The 3' ends identified using this technique cluster ~470bp downstream of the *NdhF* stop codon (Figure 17). This long 3' UTR sequence is indeed present in the dominating band in RNA gel blot analysis, as a probe detects this RNA species which starts at position 408 downstream of the stop codon (Ruwe 2010). The absence of this specific PCR product in *cp31a* mutants indicates that processing at the 3' end is dependent on CP31A.

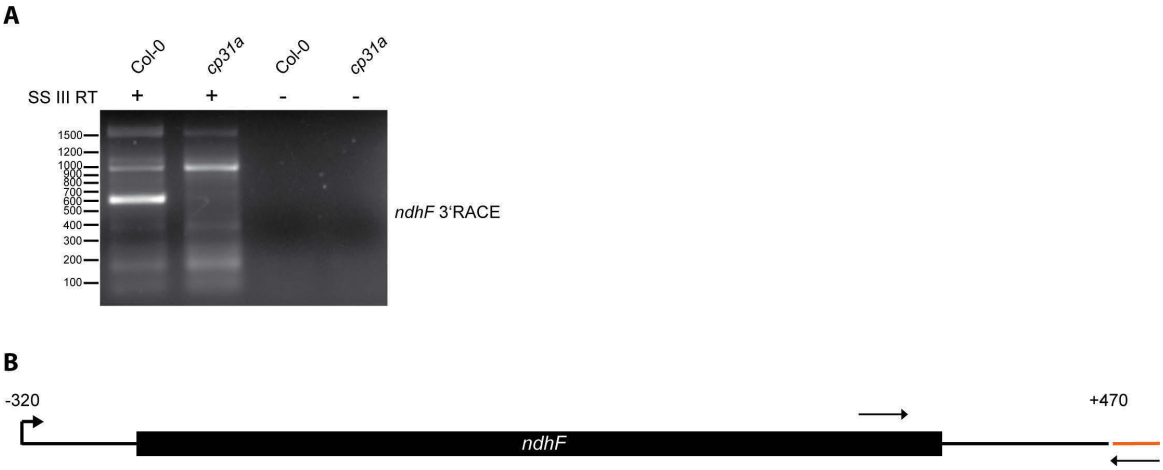


Figure 16: Identification of *ndhF* mRNA 3' ends in WT and *cp31a* mutants. (A) Total RNA from WT and *cp31a*-1 mutant tissue was ligated to an RNA oligonucleotide and reverse transcribed using an adapter-specific primer. PCR was performed using a gene-specific and an adapter-specific primer. PCR products were separated by agarose gel electrophoresis, gel-purified, and cloned (Figure 17). (B) Schematic depiction of the *ndhF* mRNA. The major 5' end was determined to be primary and strongly dependent on SIG4 (Favory et al. 2005). The position of the 3' end dependent on CP31A is 470nt downstream of the *ndhF* stop codon. The primers used for PCR amplification are shown as black arrows (not to scale). The short oligonucleotide ligated to 3' ends is shown in orange.

2.2.2 Small RNAs at the *ndhF* 3' end are reduced but not absent in *cp31a*

Processing at 3' ends is common in chloroplasts and mitochondria where transcript 3' ends are generally created post-transcriptionally (reviewed in Hammani and Giege 2014, reviewed in Stern et al. 2010). Either stable RNA structures or RBPs are needed as blocks against the action of exonucleases in the chloroplast (reviewed in Barkan 2011). The sequence upstream of the mature *ndhF* 3' is not predicted to form a stable stem-loop (Ruwe 2010). Binding sites of RBPs can accumulate as small RNAs and often do overlap with processing sites (Ruwe and Schmitz-Linneweber 2012, Zhelyazkova et al. 2012a), (2.1). Two small RNAs were identified in the region around the processed 3' end of *ndhF* (Figure 17, C79 and C80 in Supplementary Table 2). One overlaps with the *ndhF* 3' end as determined by RACE (Figure 16, Figure 17).

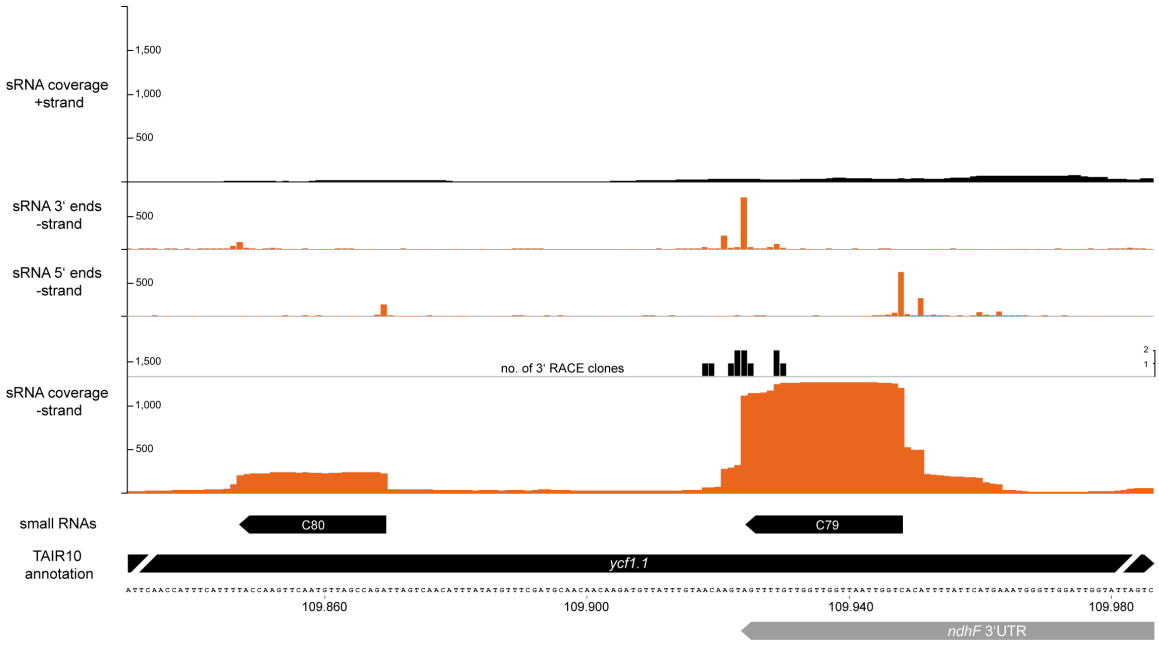


Figure 17: Transcript 3' ends of *ndhF* overlap with a small RNA. The genomic region around the identified 3' ends of *ndhF* is shown. Numbers of clones obtained by 3' RACE analysis (Figure 16) are shown as a bar graph above the small RNA coverage for the negative strand, which is shown in orange. 5' and 3' ends of small RNAs aligning with the chloroplast genome are shown in orange bars. The short *ycf1.1* gene which only encodes the N-terminal part of Tic214 is shown as a black arrow. The *ndhF* 3' UTR is indicated by a gray arrow.

Potentially CP31A binds to a sequence within the small RNA and blocks exonucleases, similar as shown for PPR10 (Pfalz et al. 2009, Prikryl et al. 2011). To test this hypothesis, an RNase protection assay was performed to analyze the abundance of small RNA species in *cp31a* mutants. A mutant of a closely related RBP, CP29A and a *cp29a*/*cp31a* double mutant were included in the analysis as well (Kupsch et al. 2012). The ends of the small RNA identified in this region are not particularly sharp as judged from the small RNA profile (Figure 17). This leads to protected fragments of slightly different sizes between 20 and 33nt. Similarly the 3' ends of *ndhF* mRNAs are dispersed over about ten base pairs (Figure 17). Many bands in Figure 18A are therefore not allocatable to a specific RNA species (small RNA or mRNA). However, species with a length below 28nt likely represent small RNAs, while species above 33nt likely represent mRNAs. Both, bands which represent mRNAs and small RNAs are slightly reduced in *cp29a* mutants. In *cp31a* mutants all bands are drastically reduced but neither small RNAs nor longer forms are absent. Double mutants accumulate even less of all RNA species (Figure 18A).

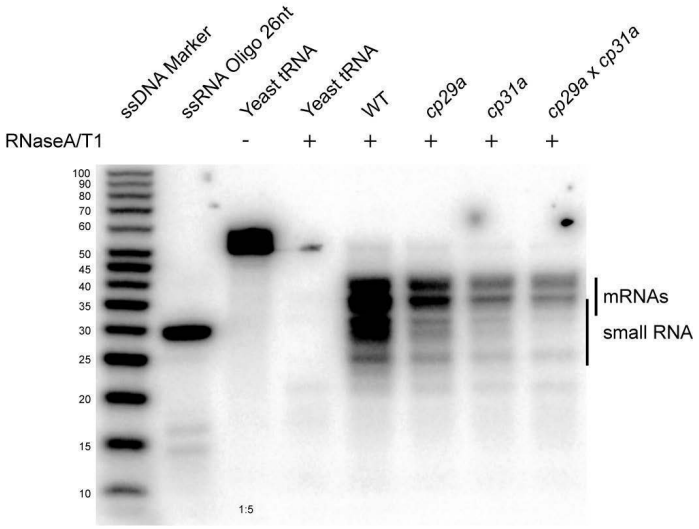


Figure 18: A small RNA at the *ndhF* 3' end is reduced in *cp31a* mutants. (A) RNase protection assay to identify small RNA accumulation in the WT, *cp29a*-1, *cp31a*-3 and *cp29a*-1x*cp31a*-3 mutants. 5µg total RNA was hybridized with a radiolabeled antisense RNA and digested with a mixture of RNase A and T1. Protected fragments were separated on denaturing polyacrylamide gels. A single-stranded DNA ladder and an RNA oligo were end-labeled and serve as size markers. Hybridization with yeast RNA controls for probe integrity during the experiment (-RNase, 1:5 dilution) and self-protection of the probe (+RNase). Fragments that originate likely from mRNAs or small RNAs are indicated.

Even though the small RNA, which coincides with the mature 3' end of *ndhF,* is strongly reduced in the *cp31a* mutant it cannot be concluded that the presence of CP31A is a requirement for its accumulation. This situation is therefore different from the cases described earlier where the absence of a PPR or HAT protein was accompanied by a complete absence of a small RNA (2.1.3).

2.2.3 Antisense transcripts of *ycf1* are dependent on CP31A

The largest proportion of the *ndhF* gene is located in the small single copy region of the *Arabidopsis* chloroplast genome. However, the last 12 amino acids of the NdhF protein and the entire 3' UTR are encoded in the inverted repeat region A (IR-A, Figure 19A). Accordingly the 3' UTR sequence is present in an additional copy in the inverted repeat B (IR-B). If RNA is expressed from this second copy its accumulation would likewise be dependent on CP31A. This hypothesis is supported by the initial finding that a strand-specific 3' UTR probe for *ndhF* detects additional bands other than the full-length *ndhF* mRNA (bands 1-6 in Figure 19B), (Ruwe 2010). Strand-specific RNA gel blot analyses were performed to elucidate the origin of the additional bands in the WT and *cp31a* mutants (Figure 19B). Probes used in these analyses are located at both border regions of the small single

copy region and inverted repeat regions A and B (Figure 19A). A probe located in the *ndhF* coding region gave rise to four distinct RNA species (1, 3, 5, and 6 in Figure 19B) which are all not detectable or nearly absent in *cp31a* mutants. Only band 1 has a size bigger than 2,000nt and thus can contain the entire open reading frame of *ndhF* (Figure 16). This band likely resembles the *ndhF* mRNA. Bands 2 and 4 are detected with a probe in sense with the open reading frame *ycf1*, thus providing evidence that these represent antisense transcripts to *ycf1*. The *ycf1* gene encodes a core subunit (Tic214) of the translocon at the inner envelope membrane (Kikuchi et al. 2013). Two short transcripts antisense to *ycf1,* band 7 and 8, are increased in abundance in *cp31a* mutants. Both are not detected with a probe in the inverted repeat region showing them having a different 3' end (Figure 19).

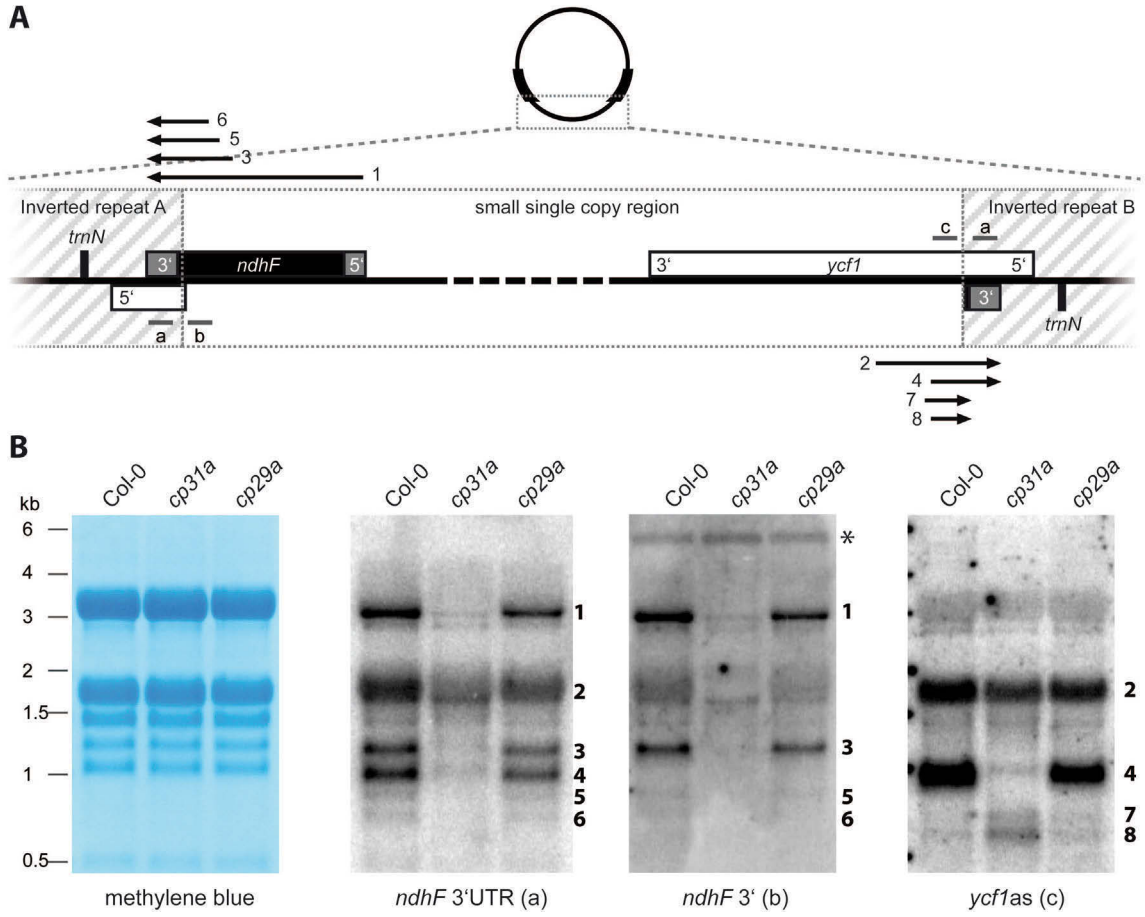


Figure 19: CP31A dependent accumulation of *ndhF* and *ycf1*as transcripts. (A) Genomic map of the two borders between small single copy region and inverted repeats A and B. Genes are indicated as black boxes. The untranslated sequences of the *ndhF* mRNA are indicated as gray boxes. Genes above the line are transcribed from left to right; genes below the line from right to left. Strand-specific probes used in RNA gel blots (B) are indicated by bars and labeled a-c. Transcripts detected are numbered corresponding to bands detected in (B). (B) RNA gel blots using strand-specific probes shown in (A). 5µg total RNA was separated in denaturing formaldehyde agarose

gels and transferred to nylon membranes. Methylene blue staining of membranes is shown as a loading control. Probe a is located in the inverted repeat region and detects *ndhF* transcripts and transcripts antisense to *ycf1*. Probes b and c are located in the small single copy region and detect *ndhF* and *ycf1*as transcripts selectively. The asterisk marks a signal for the tricistronic transcript *psaA-psaB-rps14* from a preceding hybridization.

To test whether the similarity in dependence on *cp31a* is reflected in similar 3' end processing for *ycf1*as transcripts, a 3' RACE analysis was performed in WT and *cp31a*. Figure 20 shows the results of this *ycf1*as RACE experiment. The dominant band in WT samples at around 600bp is strongly reduced in the T-DNA insertion line *cp31a*-1. The regions were gel-excised, cloned and subsequently sequenced. The positions of *ycf1*as 3' ends from several clones obtained from WT and the *cp31a* mutant are shown in Figure 20B. The 3' ends of *ycf1*as and *ndhF* transcripts are found at very similar positions, clustered at the 3' end of the small RNA in the WT. 3' ends from *ycf1*as transcripts in the *cp31a* mutant are more dispersed (Figure 20B).

In conclusion, a number of transcripts antisense to *ycf1* and transcripts in the 3' end of the *ndhF* gene share the same 3' end as the *ndhF* mRNA. All transcripts which share this 3' end are strongly reduced in plants where CP31A is not present.

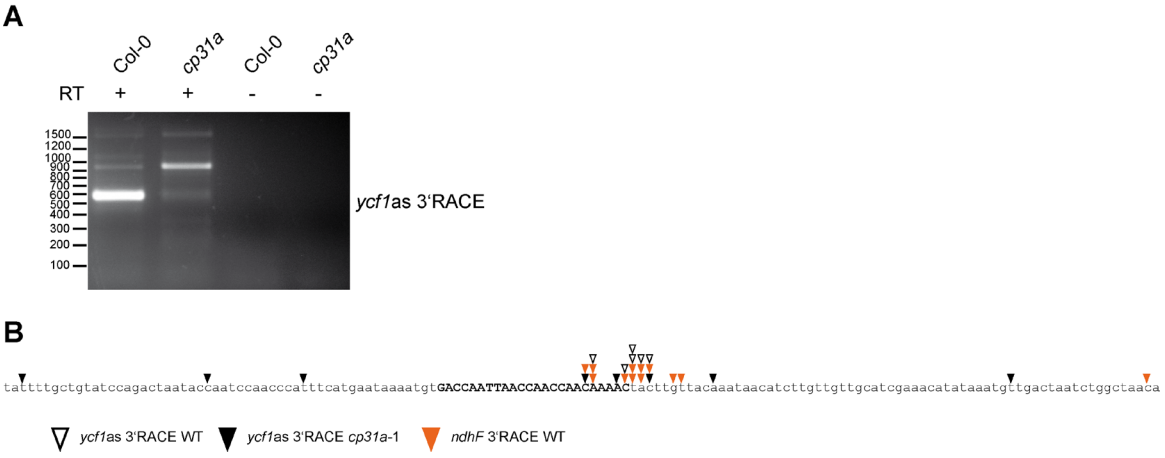


Figure 20: Identification of *ycf1*as transcript 3' ends in WT and *cp31a* mutants. (A) Total RNA from WT and *cp31a*-1 mutant tissue was ligated to an RNA oligonucleotide and reverse transcribed using a primer complementary to this adapter. PCR amplification was performed using a gene-specific and an adapter-specific primer. PCR products were separated by agarose gel electrophoresis, gel-purified and cloned. (B) Analysis of clones from 3' end mappings. The sequence of the small RNA identified at the *ndhF* 3' end is shown in uppercase bold letters. The 3' ends from *ycf1*as transcripts in WT samples are shown in open triangles, those from the *cp31a* mutant in closed triangles. For comparison the 3' ends for *ndhF* transcripts from WT samples are shown in orange triangles.

2.3 Identification of novel plastid RNA-editing sites in *Arabidopsis*

A peculiarity in organellar gene expression in land plants is the modification of genetic information on the level of RNA. RNA editing in higher plants changes cytidines to uridines by deamination.

2.3.1 Quantification of RNA editing by RNA-Seq

Current methods to quantify RNA-editing events include bulk sequencing of RT-PCR products by regular Sanger sequencing, poisoned primer extension and high resolution melting analysis (Chateigner-Boutin and Small 2007, Driscoll et al. 1989). Aforementioned methods all represent targeted approaches. Next-generation sequencing of cDNAs allows both targeted and untargeted analysis of RNA-editing events (Bentolila et al. 2013, Li et al. 2009b). For this thesis, a published strand-specific RNA-Seq library (Hotto et al. 2011) was reanalyzed to provide the first plastome-wide view of RNA editing in the model plant *Arabidopsis thaliana*. This analysis includes the quantification of RNA editing at known sites in a strand-specific manner and the identification of so far overlooked editing events.

The dataset used consists of two cDNA libraries from WT (ecotype Col-0) and two datasets from a mutant lacking chloroplast polynucleotide phosphorylase (PNPase) due to a T-DNA insertion (Hotto et al. 2011). The cDNA libraries originated from rRNA depleted total RNA, thus containing chloroplast, nuclear and mitochondrial transcripts. Reads were mapped against a modified plastid genome sequence, where positions of known editing sites were changed from C to Y to allow equal mapping of sequences from edited and unedited transcripts.

A total of 18,600,502 reads from both WT datasets was mapped to the chloroplast genome, corresponding to 37% of the reads after adapter and quality trimming. Chloroplast transcripts differ substantially in their abundance (Legen et al. 2002). Therefore, read depth at editing sites shows strong variation (Table 3). The only known editing site in *rpoC1* is represented by only 13 reads in the two combined WT samples. Three additional sites are covered with less than 50 reads in the combined WT samples (shaded in gray in Table 3). In contrast, six sites show coverage of over 1000 reads. The editing extent at sites under investigation varied between 25% and nearly 100%, with the vast majority above 80%. For all editing sites, unedited transcripts were detected (Table 3). Looking at the two WT datasets individually, the editing extend measured was reproducible (low SD) when coverage was reasonable high, i.e. more than 50 reads per sample (Table 3).

Table 3: RNA-editing extend as determined by analysis of RNA-Seq datasets. Occurrence of C or U at known editing sites in two independent WT libraries was investigated. The percentage of RNA editing is shown for the two datasets as well as the average and standard deviation for both. For comparison, data for the WT using deep sequencing of RT-PCR products taken from (Bentolila et al. 2013) is shown. Editing sites with low coverage are shaded in gray.

	WT 1			WT 2			WT		Bentolila et al.
	C	U	%	C	U	%	AVG	STDEV	
***matK*-2931**	0	52	100%	8	48	86%	93%	10%	89%
***atpF*-12707**	70	1593	96%	58	1003	95%	95%	1%	98%
***rpoC1*-21806**	5	2	29%	6	0	0%	14%	20%	16%
***rpoB*-23898**	19	73	79%	5	64	93%	86%	9%	93%
***rpoB*-25779**	3	21	88%	3	16	84%	86%	2%	94%
***rpoB*-25992**	3	15	83%	0	29	100%	92%	12%	92%
***psbZ*-35800**	12	237	95%	9	192	96%	95%	0%	94%
***rps14*-37092**	192	2925	94%	112	1627	94%	94%	0%	88%
***rps14*-37161**	174	4365	96%	103	2673	96%	96%	0%	96%
***accD*-57868**	3	541	99%	5	323	98%	99%	1%	99%
***accD*-58642**	1	11	92%	3	8	73%	82%	13%	73%
***psbF*-63985**	24	1129	98%	17	710	98%	98%	0%	99%
***psbE*-64109**	18	6745	100%	8	4765	100%	100%	0%	100%
***petL*-65716**	3	33	92%	7	30	81%	86%	7%	94%
***rps12*-69553**	78	26	25%	62	25	29%	27%	3%	28%
***clpP*-69942**	51	215	81%	34	136	80%	80%	1%	97%
***rpoA*-78691**	45	383	89%	24	290	92%	91%	2%	83%
***rpl23*-86055**	317	925	74%	233	683	75%	75%	0%	83%
***ndhB*-94999**	4	74	95%	7	94	93%	94%	1%	99%
***ndhB*-95225**	2	175	99%	1	119	99%	99%	0%	100%
***ndhB*-95608**	12	47	80%	8	31	79%	80%	0%	98%
***ndhB*-95644**	10	51	84%	18	71	80%	82%	3%	98%
***ndhB*-95650**	7	51	88%	17	74	81%	85%	5%	99%
***ndhB*-96419**	6	137	96%	19	161	89%	93%	4%	99%
***ndhB*-96579**	6	73	92%	8	59	88%	90%	3%	98%
***ndhB*-96698**	13	66	84%	20	87	81%	82%	2%	98%
***ndhB*-97016**	2	84	98%	6	79	93%	95%	3%	99%
***ndhF*-112349**	5	99	95%	2	67	97%	96%	1%	99%
***ndhD*-116281**	15	244	94%	19	155	89%	92%	4%	86%
***ndhD*-116290**	19	191	91%	19	140	88%	90%	2%	86%
***ndhD*-116494**	0	44	100%	7	52	88%	94%	8%	92%
***ndhD*-116785**	1	84	99%	3	110	97%	98%	1%	98%
***ndhD*-117166**	42	31	42%	34	32	48%	45%	4%	44%
***ndhG*-118858**	27	150	85%	28	152	84%	85%	0%	81%
total	1189	20892	95%	913	14105	94%	94%	0%	

Using a targeted approach, Bentolila and colleagues quantified the editing extend of all 34 known editing sites in *Arabidopsis thaliana* chloroplasts by massive parallel sequencing of RT-PCR products (Bentolila et al. 2013). When comparing the two strand-specific datasets that were obtained by quite different protocols, very similar results were

obtained, with the highest deviation being 17% at the only known *Arabidopsis* editing-site in the *clpP* gene.

2.3.2 Identification of undiscovered RNA-editing events by RNA-Seq

2.3.2.1 Identification of potential DNA/RNA conflicts

Most analysis of RNA editing in chloroplast transcripts focused on coding regions, mostly due to lack of high-throughput methods. With next-generation sequencing it is possible to investigate DNA/RNA inconsistencies in a whole transcriptome with high sensitivity. For the detection of these inconsistencies an algorithm for the identification of single-nucleotide polymorphisms (SNPs) was used (4.2.20). Two cDNA libraries derived from *pnp* mutants were included in the analysis, since more non-coding regions accumulate when the polynucleotide phosphorylase, a major 3'→5' exonuclease in chloroplasts, is absent in chloroplasts (Germain et al. 2011, Hotto et al. 2011, Walter et al. 2002). A "SNP" was called when a conversion was found in at least 3% of reads at a given position. In addition, this conversion had to be present in both WT or both *pnp* replicates respectively. SNPs due to mapping artifacts of nuclear sequences, identified by BLAST searches, were removed manually as were SNPs in polymeric tracks.

Table 4 shows that all possible nucleotide conversions were detected, even though many with only few occurrences. Most SNPs were found in ribosomal and transfer RNAs, which are expected to be highly modified (Karcher and Bock 2009, Majeran et al. 2012, Tokuhisa et al. 1998). The exact positions of all identified SNPs can be found in supplemental dataset 3 in Ruwe et al. (2013). Noteworthy, a high occurrence of C to U mismatches in non-tRNA/rRNA regions was found (Table 4). In *pnp* datasets two A→G and one A→C mismatch were found outside of rRNAs and tRNAs (Table 4).

Table 4: DNA/RNA inconsistencies found in RNA-Seq datasets. All possible conversions are listed. Occurrences present in both replicates in the WT or the *pnp1*-1 mutant with a frequency above 3% and coverage of greater than ten are listed.

		WT			*pnp1*-1		
genomic	cDNA	rRNA	tRNA	other RNAs (incl. mRNA)	rRNA	tRNA	other RNAs (incl. mRNA)
A	C	1	-	-	3	-	1
	G	7	1	-	7	1	2
	U	2	1	-	-	-	-
C	A	3	-	-	4	-	-
	G	2	-	-	2	-	-
	U	11	3	7	6	-	3
G	A	6	3	-	4	3	-
	C	2	-	-	2	-	-
	U	1	3	-	2	6	-
U	A	5	2	-	4	2	-
	C	5	2	-	3	2	-
	G	3	-	-	3	1	-

2.3.2.2 Novel C→U editing events show low conversion rates

A total of ten novel C→U conversions were detected outside of tRNA and rRNA coding regions. Seven of these were detected in the WT and three exclusively in *pnp* mutants (Table 5). The three sites only present in *pnp* mutants are found in non-coding regions and show increased coverage. Likely these regions are usually degraded in WT tissue by the PNPase. In total, three sites are found inside of open reading frames, namely *ndhB* and *ndhK*. Both genes encode subunits of the NDH complex. Seven sites are found in non-coding regions. Editing at *ndhK*-49849 and *ndhB*-96439 leads to codon changes. In both cases, a TCA codon is changed to a TTA codon resulting in serine to leucine change. Editing at *ndhB*-96457 is silent, as it changes an AUC to AUU codon both encoding isoleucine.

All of these newly identified sites exhibit low editing efficiencies between 4-26% (Table 5). To exclude that these sites arise through sequencing artifacts six sites identified in WT samples were confirmed by cloning and sequencing individual RT-PCR products. From 3-8% edited cDNA clones were identified at the respective positions, confirming they represent true RNA-editing sites (Table 5). Three site identified in *pnp* mutants were confirmed by Cleaved Amplified Polymorphic Sequence (CAPS) analysis by Benoît Castandet (Ruwe et al. 2013). All of nine sites tested were confirmed. In conclusion, these nucleotide changes described above likely arose through RNA editing and are referred to as novel RNA-editing sites hereafter.

Table 5: Analysis of novel C→U editing events in *Arabidopsis* chloroplasts. Genomic positions (NCBI: NC_000932) of all C→U inconsistencies identified in WT and *pnp* mutants are shown. The edited C and 19 upstream bases are shown which potentially represent the *cis*-element recognized by editing factors. Six out of seven editing events present in the WT were confirmed by sequencing of cDNA clones.

			WT		*pnp1-1*		cDNA Cloning	
	genome position	*cis*-element	coverage	editing	coverage	editing	coverage	editing
***atpH* 3'UTR**	13210	GTAGTTTTTTTAATTCTAT**C**	2702	4%	4254	4%	76	8%
***ycf3* Intron 2**	43350	GACTAGATATGCCTAAATA**C**	390	12%	1685	1%	38	5%
***rps4* 3'UTR**	45095	ATTTTTCCTATTCATGTAT**C**	69	10%	205	1%	35	3%
ndhK	49849	AATGATCTTTCAAATTGGT**C**	124	4%	100	0%		
ndhK-ndhJ	49209	CTTCATAAATTAGAATTAA**C**	1342	6%	864	0%	43	7%
***rps18* 3'UTR**	68453	ATTTCTACTCTACCTTCCC**C**	25	0%	721	26%		
***ycf2* as**	91535	TCATCAATATCGATATCAT**C**	2	0%	47	11%		
***ndhB* 3'UTR**	94622	CTACTTTTTACATATCTCT**C**	2	0%	324	6%		
ndhB	96439	TCACTGTAGGAATTGGGTT**C**	419	6%	597	2%	41	7%
ndhB	96457	CAATTGCGCTTATATTCAT**C**	518	5%	820	2%	41	5%
ndhB	*96419*	These are three known sites present on the PCR product for *ndhB*					*41*	98%
ndhB	*96579*						*41*	98%
ndhB	*96698*						*41*	100%

3 Discussion

In this thesis, properties of two families of RNA-binding proteins (RBPs) have been investigated. Pentatricopeptide repeat (PPR) proteins represent one of the largest protein families in land plants, with about 450 members in *Arabidopsis*. One third is predicted or experimentally verified to be plastid localized, almost all other members are imported into mitochondria (Colcombet et al. 2013, Lurin et al. 2004). PPR proteins are expressed at relatively low levels, and each PPR protein is believed to target only few RNAs (reviewed in Barkan and Small 2014, Lurin et al. 2004). In contrast, chloroplast ribonucleoproteins (cpRNPs), represented by ten members in *Arabidopsis*, are highly abundant (Nakamura et al. 2001, reviewed in Ruwe et al. 2011). CpRNPs were shown to bind multiple RNAs (Kupsch et al. 2012, Nakamura et al. 1999). Despite these differences, PPRs and cpRNPs were described to act in the same processes, including RNA stabilization and RNA editing (reviewed in Barkan and Small 2014, Kupsch et al. 2012, Nakamura et al. 2001, Tillich et al. 2009). Findings on mechanistic aspects of PPR proteins and functions of cpRNPs are discussed in the following sections.

3.1 Small RNAs predicts binding sites for RNA-binding proteins (RBPs)

3.1.1 The origin of RBP footprints in plastids

The best studied example of a PPR protein in plants is PPR10. Two target sites are known in maize, and crystal structures of RNA-free and RNA-bound states are available (Pfalz et al. 2009, Yin et al. 2013). However, these structures were later challenged as they show dimeric complexes, likely an artifact of high protein concentrations needed for crystallization (Barkan et al. 2012, Gully et al. 2015). A finding with utmost importance for this thesis is that a small RNA, which carries the binding site for PPR10 in the center, can be identified in small RNA databases (Pfalz et al. 2009). The small RNA was reproduced, *in vitro*, by exonucleolytic trimming of a PPR10-bound precursor RNA, resulting in an *in vitro* footprint of the RBP (Prikryl et al. 2011). *In vivo*, a similar scenario is anticipated, with endogenous exonucleases trimming precursor RNAs until they are stopped by PPR10 (Pfalz et al. 2009).

3.1.2 How many small RNAs identified represent RBP footprints?

In this thesis, a published small RNA dataset was reanalyzed to identify additional RBP footprints, generated similar as the footprint of PPR10. About 240 small RNAs have been identified by this analysis in the chloroplast of *Arabidopsis*, using an algorithm that detects a rapid increase or drop in small RNA coverage (i.e. a peak with at least one sharp end, Figure 5, 4.2.19). This algorithm will thus detect additional small RNAs beside protein footprints. Their potential origin will be discussed in the following section.

3.1.2.1 Small RNAs accumulate from structured RNAs

Prominent RNA species identified within the 240 small RNAs are tRNA fragments, which are likely generated by endonucleolytic cleavage from mature tRNAs (reviewed in Raina and Ibba 2014). A total of 30 tRNA genes are annotated for the chloroplast genome of *Arabidopsis thaliana*. 47 tRNA derived fragments overlapping these annotations were identified (Figure 6). These could represent intermediates of tRNA degradation, although recent findings support a role for tRNA fragments in regulation of gene expression as part of a stress response in different domains of life (reviewed in Raina and Ibba 2014). Evidence for a role of tRNA fragments in chloroplasts and mitochondria is missing, but in Chinese cabbage chloroplast tRNA fragments were shown to increase under heat stress conditions (Wang et al. 2011). The same study also identified rRNA fragments predominantly at 3' ends of rRNAs. Such fragments were also identified in this thesis for *Arabidopsis*. It is not clear how tRNA and rRNA fragments are stabilized, but it is possible that they remain bound and protected from nucleases in tRNA structures and ribosomes. Some small RNAs were identified antisense to tRNA genes (Figure 6), which indicates that stable structures forming in antisense orientation to tRNAs are sufficient to stabilize small RNAs. Importantly, these sequences are likely not protected by RBPs (ribosomal proteins stabilizing rRNA derived fragments might represent the exception). When tRNA and rRNA derived fragments are removed from the dataset of small chloroplast RNAs about 180 small RNAs remain. This set of 180 small RNAs represents the first plastome-wide compendium of candidates for protein-mediated protection of small RNAs. Whether the set is complete or whether in other tissues or under different conditions more small RNA will be identified needs to be determined.

Stable RNA structures are able to block exonucleases (reviewed in Stern et al. 2010, Stern and Gruissem 1987) and thus act similar as a protein cap represented by an RBP.

Figure 5 shows that small RNAs accumulate in the 3' UTR of *rbcL*, a transcript which terminates in a stable stem-loop structure (Zurawski et al. 1981). A small RNA identified by the algorithm overlaps with the predicted stem-loop structure. This indicates that the small RNA is protected by an RNA-RNA hybrid rather than an RBP. When considering small RNAs as candidates for protein binding sites, a structure prediction should thus always be performed (Ruwe and Schmitz-Linneweber 2012, Zhelyazkova et al. 2012a). Nine small RNAs resulting likely from stem-loop structures ($\Delta G < -20$kcal/mol) were identified in a different small RNA dataset (Rajagopalan et al. 2006, Ruwe and Schmitz-Linneweber 2012). At six out of these nine genomic regions small RNAs were identified in this thesis as well. In two regions, only small RNAs on the opposite strand were identified. At one position small RNAs were identified on both strands (*psbM-petN*). Likely these stem-loop structures can block exonucleases in sense and antisense transcripts, as shown in *Chlamydomonas* chloroplasts (Rott et al. 1998). Thus these stem-loop structures could be especially beneficial between convergent genes (stem-loops identified in intergenic regions of convergent genes: *psbM-petN*, *psbC-trnS*, *atpE-trnM*, *petA-psbJ*, *psbT-psbN*, *and petD-rpoA*). In contrast, on parallel oriented genes stable stem-loop structures could be deleterious stabilizing antisense transcripts.

3.1.2.2 Small RNAs that represent footprints of RBPs

The majority of the 180 small RNAs is found in non-coding regions, preferred locations for RBPs involved in intergenic processing and translation initiation (Figure 6). Similar to the footprint of PPR10, small RNAs in intergenic regions and rarely also in coding regions (*psbC* 5' end in *psbD*), do overlap with processed transcript ends as identified by transcript end mapping (Figure 8), (Ruwe and Schmitz-Linneweber 2012). This finding and the absence of specific small RNAs in mutants of RBPs supports the idea that these small RNAs are generated similar as the PPR10 footprint (Figure 9). Additional end mappings in *Arabidopsis* and barley support the frequent coincidence of small RNA accumulation with processing sites (Malik Ghulam et al. 2013, Zhelyazkova et al. 2012a). Even though predominantly a single small RNA is found per intergenic region, deviations from that rule can be observed. In the *rps7-ndhB* intergenic spacer two small RNAs can be identified and both small RNAs overlap with mapped transcript ends (Figure 8).

The identification of many overlapping transcripts indicates that the initial idea of processing by a single endonucleolytic cleavage event, an idea that resulted from imprecise

mapping of transcript ends, is outdated (reviewed in Barkan 2011). Comparing small RNA mappings obtained in this thesis with known transcript patterns of well-studied plastid operons, e.g. the *rps2*/*atpI*/*atpH*/*atpF*/*atpA* operon and the *psbB*/*psbT*/*psbH*/*petB*/*petD* operon, shows that small RNAs are present in all intergenic spacers subjected to processing (Figure 7), (reviewed in Barkan 2011, Meierhoff et al. 2003, Pfalz et al. 2009, Sane et al. 2005). Identifications of small RNAs in the majority of *Arabidopsis* intergenic spacers indicate that most intergenic processing activities in chloroplasts are due to the protection against exonucleases through binding of RBPs. Using deep next-generation sequencing of small RNAs, processing sites can thus be predicted from the accumulation of small RNAs in intergenic regions (Figure 8). In summary, this thesis shows on a transcriptome-wide level that intergenic and in part end processing in chloroplasts is achieved via the joined action of RBPs and exonucleases.

3.1.3 Which RBPs leave footprints?

A total of 154 PPR proteins have been predicted or experimentally shown to be imported into the chloroplast and about 320 in mitochondria (Colcombet et al. 2013). Members of other RBP families extend this list (reviewed in Jacobs and Kuck 2011). Taken together, there is a large potential for factors generating small RNAs, foremost PPR and PPR-like proteins.

3.1.3.1 Overlap of small RNAs with described processing sites

The model of protein-mediated protection of small RNAs has been validated for a handful of small RNAs representing footprints of PPR and related tetratricopeptide repeat (TPR)-like proteins in chloroplasts and mitochondria (footprints of: PPR10, CRP1, HCF152, MTSF1, MRL1, HCF107, and Mbb1). PPRs and PPR-like proteins are thus, based on the so far identified RBPs involved in intercistronic and end processing of organellar transcripts, the best candidates. Exceptions are PrfB3, a relative of ribosomal release factors, and CP31A. The two RBPs are involved in intergenic processing between *petB*-*petD* and the end processing of *ndhF* respectively (Kupsch et al. 2012, Stoppel et al. 2011). CP31A was shown in this thesis to be beneficial but not essential for the accumulation of a small RNA, at the 3' end of *ndhF* (Figure 18). Therefore, it can be concluded that the small RNA at the end of *ndhF* does not represent the footprint of CP31A. A potential target of PrfB3, a small RNA downstream of *petB* which overlaps with the processed 3' end of *petB*

in *Arabidopsis* and maize was shown to be the target of CRP1 (Zhelyazkova et al. 2012a). The *prfb3* mutants are still able to perform correct processing, but with strongly reduced efficiency, a scenario which mirrors the situation in *cp31a* mutants and the *ndhF* message (Stoppel et al. 2011). CP31A was shown to bind multiple RNAs with multiple interactions per mRNA (Kupsch et al. 2012). The interaction strength is likely not sufficient to block exonucleases, otherwise small RNAs would accumulate throughout messages targeted by cpRNPs. Potentially PrfB3 and CP31A act together with helical repeat proteins to stabilize the processed *petB* and *ndhF* mRNAs. Thus the only proteins known to leave *in vivo* footprints are members of the PPR and TPR-like families.

3.1.3.2 Different classes of PPR proteins leave *in vivo* footprints

PPR proteins can be divided in two classes, based on the types of repeats found in the proteins (Lurin et al. 2004). P-class PPR proteins like PPR5, PPR10, MRL1 and HCF152 have been implicated in RNA processing and stabilization, whereas PLS-class proteins are mostly implicated in RNA editing (reviewed in Barkan and Small, 2014).

PLS-DYW protein CRR2 presents an exception from that basic rule as it is implicated in the intergenic processing between *rps7* and *ndhB* that was believed to result from intrinsic endonucleolytic activity of CRR2 (Hashimoto et al. 2003, Okuda et al. 2009). In this thesis, it was shown that CRR2 leaves a footprint overlapping the processing site. In disagreement with the proposed cleavage mechanism, CRR2 dependent ends of *rps7* and *ndhB* overlap by about 24nt (Figure 8, Figure 12).

In contrast to CRR2, most PLS PPR proteins do not leave small RNA footprints at known target sites. At none of the 34 known RNA-editing sites, which are likely all recognized by PLS-class PPR proteins (reviewed in Shikanai 2015), small RNAs accumulate. The only RNA-editing site that overlaps with a small RNA is one newly identified partial RNA-editing site in the 3'UTR of *rps18* (Table 5, C40 in Supplementary Table 2). This small RNA includes the editing site and the potential *cis*-element (21nt upstream of the editing site are present in the small RNA). The small RNA overlaps the dominant transcript end of *rps18* and is thus likely a footprint of an RBP stabilizing the *rps18* mRNAs (Ruwe and Schmitz-Linneweber 2012). Potentially it is the same factor that stabilizes *rps18* and edits this site.

Most editing sites are found in coding regions and tight binding of an editing factor might interfere with translation of the open reading frame (reviewed in Barkan and Small

2014). Potentially, RNA-editing factors are counter selected against tight binding. One could speculate that even though specificity of RNA-editing factors needs to be high, the affinity might be required to be lower than for PPR proteins blocking exonucleases. However, a couple of small RNAs accumulate in coding regions, indicating that RBPs with high affinity bind there, without disturbing translation (Figure 6). Some are found in the upstream open-reading frame of overlapping or closely spaced genes (*psbD-psbC*, *ndhH-ndhA*, *rpoB-rpoC1*, and *rpl32-rpl2*). The small RNA found in the open reading frame of *psbD* overlaps with transcript 5' ends of the downstream gene *psbC*. By contrast, no incomplete *psbD* transcripts overlapping the small RNA could be detected (Figure 8). One possible explanation for this finding is that chloroplast ribosomes are able to displace this RBP from RNA that cannot be displaced by exonucleases. This would lead to a situation, where downstream *psbC* can be translated from transcripts with 5'ends defined by the unknown RBP, whereas the *psbD* ORF is only translated from dicistronic messages. The situation might be different in mitochondria, as three transcript 3' ends overlapping with small RNAs (*nad6*, *mttB*, *ccmC*) are located inside of open reading frames (Table 2), (Forner et al. 2007). The resulting non-stop mRNAs are very likely translated in cauliflower mitochondria (Raczynska et al. 2006). An interesting hypothesis is that in these open reading frames, RBPs stop ribosomes and initiate translation termination in mitochondria.

In conclusion, the accumulation of small RNAs that represent footprints of PLS-class PPR proteins might be a rare occurrence that is accomplished by changes in the protein sequence and structure that lead to tighter binding (discussed below). In addition, certainly not all P-class PPR proteins leave footprints as only a total of 7 small RNAs in intronic sequences have been identified and the number of PPR proteins involved in intron splicing is likely much higher (reviewed in Barkan and Small 2014). It seems that only PPR proteins required for processing/stability leave footprints and one possible explanation is a higher affinity of these proteins for their RNA targets.

3.1.4 Identification of additional targets of PLS-DYW protein CRR2 increases the understanding of PPR-RNA interactions

In *crr2* mutants eleven small RNAs are missing as evident from small RNA sequencing (Figure 10). Their similar length further supports that they are footprints of the same RBP, CRR2 (Figure 10, Figure 12). With the exception of RNase P, which also contains three PPR repeats, CRR2 thus represents the PPR protein with the most known RNA

targets (Gobert et al. 2010). This relatively large number of targets identified in this thesis could help to understand the mechanism underlying these protein-RNA interactions.

An alignment of the eleven target sites shows that bases in the center of the small RNAs are more similar (Figure 12). Exterior, non-conserved bases likely represent bases not bound by the PPR protein. Speculatively these exterior bases cannot be further trimmed by exonucleases and potentially represent the physical distance from the active center to the surface of the exonucleases (Germain et al. 2012).

PLS-class PPR proteins were believed to bind RNA bases in a consecutive manner, with L motifs not involved in RNA binding to reduce structural constraints (Barkan et al. 2012). In contrast, bases aligned with L motifs in the eleven small RNAs show a bias towards specific nucleotides. The third L motif is aligned with a U in nine out of eleven small RNAs. Furthermore, L motifs one and two are enriched for U and G respectively (Figure 21). In accordance, including L motifs in target predictions of PLS-class RNA-editing factors improves accuracy. This suggests that L motifs interact with RNA and provide specificity (Takenaka et al. 2013a).

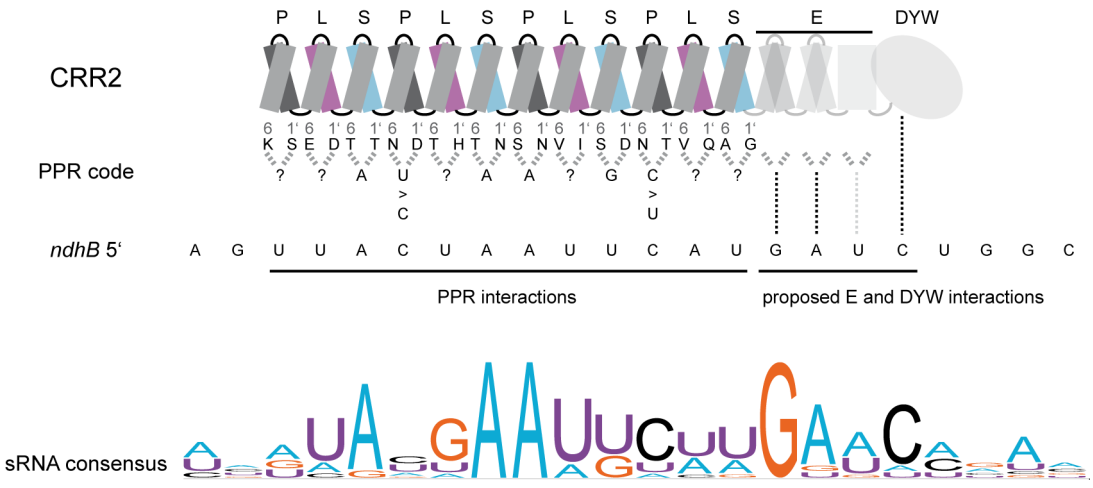


Figure 21: Involvement of C-terminal domains in CRR2-RNA interaction. The domain structure of CRR2 with four blocks of P-L-S repeats and the C-terminal E and DYW domains are shown. Amino acids 6 and 1’ involved in base recognition are indicated as are the preferred bases according to the PPR code. The main target *ndhB*5’ is aligned and is in accordance with the PPR code with the exception of the U found opposite of the third S motif. The consensus sequence of the 11 small RNAs targeted by CRR2 indicates that bases outside the PPR alignment are more conserved and potentially targeted by the E and DYW domains of CRR2.

The third S motif in CRR2 aligns exclusively with U and G in the eleven small RNAs (Figure 21). Only G is in agreement with the PPR code in S motifs, as proposed by Barkan et al. (2012). The occurrence of U in many small RNAs at this position, including

the most abundant small RNA upstream of *ndhB*, suggests that U is specifically recognized or at least tolerated at this position. Bases that align with other P and S motifs of CRR2 aside of the third S motif are in accordance with the code suggesting that the overall alignment is correct (Figure 21).

3.1.4.1 C-terminal domains in CRR2 provide specificity

Four bases 3' of the alignment with PPR motifs show similarity in the eleven target sites of CRR2 (Figure 21). Suggesting the alignment is correct, it is very likely that the specificity of protein-RNA interaction is also determined by additional domains other than PPR motifs. The fourth base, a consensus C, is found at the position where the C to be edited is found in alignments between editing factors and their target sites (Barkan et al. 2012, Takenaka et al. 2013a, Yagi et al. 2013). CRR2, as many other PLS-class proteins, carries C-terminal extensions, namely an E and DYW domain. Both domains are frequently found in PPR proteins implied in RNA editing (reviewed in Shikanai 2015). It has been hypothesized, that the DYW domain carries the catalytic activity, based on similarities with cytidine deaminases (Salone et al. 2007). Recently the DYW domain has been implicated in providing specific recognition of the C to be edited (Okuda et al. 2014).

Mutational analysis of the DYW and also the E domain of CRR2 support the idea that the DYW domain and the E domain provide specificity *in vitro* (Peter Kindgren, personal communication). The current working hypothesis is that the E domain of CRR2 specifically recognizes the GA found in most sites (Figure 21). The E domain resembles highly degenerated PPR repeats, thus RNA-binding activity of the E domain is conceivable (Okuda et al. 2007, reviewed in Takenaka 2014, Wagoner et al. 2015, Yagi et al. 2013). Potentially, other PLS-class proteins rely on this interaction as well, but for some factors the E domain has been demonstrated to be dispensable for RNA-binding *in vitro* (Okuda et al. 2014). An additional candidate for specific interaction of the E domain and target RNAs is CRR28. Two target sites are known, but CRR28 is not associated with a high score with this genetically identified targets using the PPR repeats only and the described PPR code (Barkan et al. 2012, Takenaka et al. 2013a, Yagi et al. 2013), (Supplementary Table 1). The two target sites of CRR28, *ndhB*-96698 and *ndhD*-116290, both carry a CU at position -3 to -2 with respect to the edited C that could be recognized by the E domain of CRR28. Indeed, recombinant CRR28 binds RNAs, where bases from -3 to -1 are deleted, with reduced affinity (Okuda et al. 2014).

In summary, in this thesis a number of previously unknown RNA targets have been identified for CRR2. An alignment of the protein with the RNA targets suggests that the E and DYW domain provide specificity for RNA recognition. This finding could hold true also for other PLS-class PPR proteins.

3.1.4.2 CRR2 an editing factor that lost its editing activity?

For none of the sites recognized by CRR2, C to U conversion was found (Ruwe et al. 2013). In an alignment of the DYW domain of CRR2 with DYW domains of editing factors CRR22, CRR28, and YS1, deviations from the DYW consensus sequence at highly conserved positions were identified for CRR2 (Okuda et al. 2009). CRR2 might thus represent an editing factor that lost the ability to deaminate or recruit the deamination activity to its targets sites, but targets with a C at the position to be edited are still preferentially recognized, likely by the DYW domain. Investigations of chimeric proteins of CRR2 and DYW domains of editing factors could help to understand target specificity of PLS-class proteins and could shed light on the editing mechanism in general.

Many CRR2 binding sites identified by small RNA sequencing are potentially off-targets. The ten additional small RNAs, other than *ndhB* 5', show weaker coverage by a factor of at least 40 in small RNA libraries. Even though many small RNA sequencing protocols are not strictly quantitative (Hafner et al. 2011), this finding still indicates that the sequence upstream of *ndhB* is the prime target of CRR2. A CRR2 dependent small RNA downstream of *ycf2* is a good candidate for an overlap with the mature *ycf2* 3' end. However, in a 3' RACE analysis the mature 3' end was detected further downstream, likely overlapping a second small RNA 150nt away (data not shown, C132 in Supplementary Table 2). In support of this finding, transcripts of *ycf2* detected in RNA gel blot analysis did not show any alteration in *crr2* mutants (Figure 13B). The high number of off-targets of CRR2 might indicate that P-class proteins are better suited to fulfill the job to stabilize and increase translation of plastid transcripts, since they are acting with higher specificity than PLS-class proteins.

3.1.5 Using small RNA accumulations to identify RBP targets

The set of 180 small RNAs described above could serve as a template for the bioinformatic prediction of target sites for PPR proteins, similar as performed for RNA-editing factors using the *cis*-elements found upstream of editing sites (Barkan et al. 2012, Takenaka

et al. 2013a, Yagi et al. 2013). Reducing the sequences to search against, by using the small RNAs identified in contrast to a the complete genome sequence, might be especially helpful, because alignments of P-class PPR proteins with RNA targets is challenging. Investigations so far concluded that neighboring PPR repeats in P-class PPR proteins not necessarily bind contiguous RNA bases (reviewed in Barkan and Small 2014). In addition, alignments of PPR10 and CRP1 with known targets suggest that the mode of RNA recognition can vary between different RNA targets of a single PPR protein (Barkan et al. 2012). More specific, varying numbers of nucleotides can be tolerated between two stretches of RNA bases specifically recognized (Barkan et al. 2012).

More direct as the bioinformatic prediction is the identification of RNA targets of RBPs involved in intercistronic and end processing by sequencing. As shown in this thesis, this is approachable by small RNA sequencing from total RNA isolated from mutant material. The comparison of small RNA accumulation between mutants of RBPs and the WT provides a rapid method to map the exact binding sites of a protein in a transcriptome-wide manner (Figure 10). For RBPs that do not leave protein footprints *in vivo*, modifications of the protocol could discover nuclease sensitive sites in mutants of RNA-binding proteins. Digestion of extracts with endonucleases or less processive exonucleases would allow the identification of footprints with an affinity that is too low to block endogenous exonucleases (Liu et al. 2013a, Silverman et al. 2014).

3.1.5.1 PPR-SMR protein SOT1 is required for ribosomal RNA maturation

Analyses on *crr2* mutants described above and *sot1* mutants highlight the potential of small RNA sequencing for the discovery of PPR-RNA interaction sites. In the *sot1*-2 mutant, absence of three small RNAs with similar sequence was discovered (Figure 11). A small RNA upstream of the *rrn23* gene is highly abundant and the absence of the small RNA in *sot1* mutants is paralleled by ribosomal RNA processing defects (Dr. Kate Howell, personal communication). The small RNA upstream of *rrn23* does overlap with the 5' end of a precursor of the 23S ribosomal RNA (Bollenbach et al. 2005). 5' RACE analysis showed that this processed 5' end is absent in *sot1*-2 mutants (Dr. Kate Howell, personal communication and Supplementary Figure 1). This finding suggests that SOT1 stabilizes the precursor and allows proper ribosomal RNA maturation and ribosome biogenesis. Potentially, SOT1 protects the precursor against the 5'→3' exonucleolytic activity of RNase J. Using an oligonucleotide probe to detect the small RNA upstream of *rrn23*, additional

hybridization signals were obtained (Figure 11). An abundant RNA species of about 75nt is missing in *sot1* mutants. The hybridization signal with 75nt in RNA gel blots could represent the entire precursor sequence, from the SOT1 binding site to the mature 5' end of the 23S rRNA. The accumulation of this fragment could indicate that 5' maturation of 23S rRNA is performed by an endonucleolytic rather than a trimming activity. This model was recently verified, and two paralogous genes encoding double-strand-specific Mini-III endoribonucleases were shown to be required for this cleavage event (Hotto et al. 2015). Figure 22 shows a model that could explain how SOT1 binding influences 23S maturation. By impeding RNase J progression, SOT1 allows cleavage by Mini-III that recognize the sequence in a double strand that likely forms from the 5' part of 23S and 3' part of 4.5S RNA (Massenet et al. 1987).

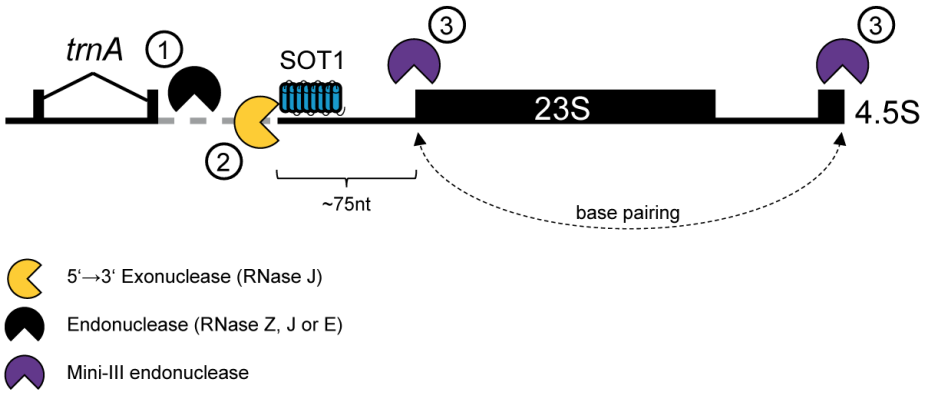


Figure 22: Model for the 5' maturation of plastid 23S rRNA. The endonuclease activity of RNase Z maturing *trnA*-UGC or other endonuclease activities create entrance sites for RNase J upstream of the SOT1 binding site (1). When SOT1 is present, RNase J is blocked about 75nt upstream of the 5' processing site (2), allowing proper 5' maturation of 23S and 3' maturation of 4.5S by Mini-III (3). If SOT1 is missing, RNase J can progress into the 23S rRNA which is accompanied by rRNA processing defects.

A second SOT1 dependent small RNA was found upstream of *ndhA*. The small RNA overlaps with a primary transcript end as determined by 5' RACE (Supplementary Figure 1). Furthermore, sequencing of tobacco small RNAs after treatment with a phosphorylation sensitive 5'→3' exonuclease suggests that many small RNAs carry a triphosphate at this position (Gongwei Wang, personal communication). Likely SOT1 is not required for stabilization of transcripts but could be required for translation of *ndhA*, potentially by structure remodeling around the start codon as shown for PPR10 and HCF107 *in vitro* (Hammani et al. 2012, Prikryl et al. 2011). The 5' UTR of *ndhA* has been speculated to be a target for

PPR protein PGR3 (Cai et al. 2011). It will be interesting to see whether both proteins bind in the relatively small 5'UTR of only 67nt.

A third small RNA was identified antisense to *rpoA*. This small RNA is represented by fewer reads in small RNA sequencing and could represent an off-target of SOT1. Detection of low abundant off-targets for SOT1 and CRR2 indicates that the coverage obtained in these experiments is sufficient.

3.1.6 Mitochondrial small RNAs

3.1.6.1 Small RNAs at 3' ends of mitochondrial transcripts implicate PPR proteins in stabilization of mitochondrial transcripts

Mitochondrial-encoded genes in *Arabidopsis* are usually separated by several kb of genomic sequence. In turn, polycistronic mRNAs are rare in mitochondria. Nevertheless, processing of 5' and 3' ends is a common feature of mitochondrial RNA metabolism (reviewed in Hammani and Giege 2014). While transcripts of mitochondrial genes often show several processed 5' ends, usually only single 3' ends are observed (Forner et al. 2007). In addition, positions of 5' ends are not very well conserved even between different *Arabidopsis* ecotypes (Forner et al. 2008). The generation of both 5' and 3' ends has been assumed to rely on specific RNA folds, with similarity to tRNA structures. These structures have been speculated to be recognized by enzymes that cleave precursor tRNAs (Forner et al. 2007). In the last years several P-class PPR proteins have been described to support processing of individual mitochondrial transcript ends (reviewed in Binder et al. 2013). PPR proteins involved in 5' processing were predicted to bind upstream of the processing sites and are believed to facilitate endonucleolytic cleavage, potentially by stabilizing beneficial RNA structures. Both 5' ends of mature transcripts and 3' ends of leader sequences could be mapped in the same region, which supports a model of endonucleolytic cleavage (Jonietz et al. 2011). This finding points to a difference between 5' processing in mitochondria and chloroplasts of land plants, where the roadblock mechanism seems to be more prominent. In line with this finding, only few small RNAs identified in this thesis overlap with mitochondrial 5' ends (Table 2). Maybe the essential difference is the presence of RNase J in chloroplasts, while a 5'→3' exonucleolytic activity seems absent in mitochondria (Sharwood et al. 2011). The lack of a 5'→3' exonucleolytic activity poses a problem for the generation of mitochondrial protein footprints in general. Indeed, small RNAs that were detected at 3' ends of mitochondrial transcripts showed broad distribution of 5' ends

(Figure 15). This broad distribution of 5' ends could be explained by stochastic endonucleolytic generation of small RNA 5' ends. The 3' ends of small RNAs are sharp and likely shaped by the action of exonucleases like the PNPase and RNR1 (Perrin et al. 2004).

So far, only a single PPR protein was shown to be required for the 3' processing and the stabilization of an individual mitochondrial mRNA. MTSF1 binds in the 3' UTR of *nad4* and likely acts similar as described for plastid PPR proteins as a roadblock against 3'→5' exonucleases (Haili et al. 2013). The *nad4* mRNA 3' end and in general 70% of mapped mitochondrial 3' ends are associated with a small RNA (Table 2). This large number of small RNAs at transcript 3' ends predicts that the majority of mitochondrial mRNAs is stabilized by the binding of RBPs. Especially in mitochondria, where mRNA levels were shown to be adjusted by the 3'→5' exonuclease PNPase (Giege et al. 2000, Holec et al. 2006), a rate-limiting role for PPR proteins in the determination of transcript levels can be anticipated. If most transcripts in mitochondria are stabilized by PPR proteins as predicted by the strong overlap of small RNA and transcript 3' ends (Table 2), changing the level of a PPR protein in the background of access transcription could determine the number of transcripts accumulating. This hypothesis could be tested by artificially overexpressing a specific PPR protein and measuring the abundance of the target transcript.

3.1.6.2 24nt long small RNAs likely originate from NUMTs

Small RNAs that map to the mitochondrial genome showed a bias towards sequences with 24nt length (Figure 4). In plants, accumulation of 24nt long siRNAs coincide with heterochromatic regions in the nuclear genome (Zhang et al. 2006). The nuclear genome of *Arabidopsis* contains a large insertion of mitochondrial DNA in the centromeric region of chromosome 2 (Lin et al. 1999). The sequence divergence between this insertion and the mitochondrial genome is very low (< 4%), indicating the insertion was a recent event (Michalovova et al. 2013). Due to this low divergence nuclear siRNAs can often map equally well to the mitochondrial genome. Centromeric regions are in general heterochromatic and also associated with 24nt long siRNAs (Kasschau et al. 2007). It has to be kept in mind that small RNAs described in this thesis can originate either from mitochondria or nuclear mitochondrial DNA (NUMTs). Actually, 24nt long small RNAs could serve as a tool to identify NUMTs and potentially also nuclear plastid DNA (NUPTs). Such an approach has recently been applied to annotate transposable elements, which are similarly associated with 24nt long small RNAs (El Baidouri et al. 2015). Importantly, small RNAs

that overlap processing sites did not show a uniform length of 24nt and are thus more likely to result from mitochondria. A final proof could be obtained from small RNA sequencing from purified mitochondria.

3.1.7 Small RNAs in organelles: Just degradation products?

It can be assumed that many small RNAs identified in this thesis represent footprints of PPR and PPR-like proteins. While the generation by the roadblock mechanism (Figure 2) is relatively clear, knowledge about potential functions of these small RNAs is lacking. Small RNAs have been described in pro- and eukaryotic systems as regulators of gene expression. In eukaryotes, the small RNA repertoire includes miRNAs, siRNAs and piRNAs. All of these small RNAs interact with Argonaute proteins and target mostly RNA to influence stability and translation by imperfect base pairing (reviewed in Meister 2013). In bacteria small RNAs between 50-300nt are involved in gene expression often by imperfect base pairing with RNA targets and influence translation and RNA stability (reviewed in Bobrovskyy and Vanderpool 2013). Also cyanobacteria, the ancestors of plastids, use small RNAs to regulate their gene expression (Georg et al. 2014, Steglich et al. 2008).

To be able to act as a riboregulator through base pairing, small RNAs that represent footprints of PPR and PPR-like proteins need to detach from the RBP. The sequence-specific recognition of PPR proteins would, based on models and crystal structures, interfere with additional base pairing of the target small RNAs (Fujii et al. 2011, Gully et al. 2015, Yin et al. 2013). Co-immunoprecipitation with PPR10 indicated that the majority of a small RNA upstream of *atpH* is bound by its cognate RBP and could thus not act as a riboregulator (Figure 14). Still a minor fraction could be protein unbound, available to base-pair. A thorough quantification of precipitated protein, that failed so far, would allow the estimation of the size of this free pool. Whether other small RNAs are equally well bound by their cognate RBP needs experimental proof, preferentially on a genome-wide level. Biochemical separation of protein-RNA complexes and free small RNAs should be possible by differences in size, density, accessibility to ribonucleases or affinity for certain matrices. Combined with small RNA sequencing, these purifications should allow the estimation of protein bound and unbound small RNA pools. If free small RNAs exist and can persist in the organelles for sufficient time, regulatory functions in *trans* or in *cis*, on antisense transcripts, are conceivable. Overexpression of small RNAs in tobacco, or other species susceptible for plastid transformation, could be used to identify targets of specific small RNAs.

The relative abundance of the small RNAs together with the finding that at least PPR10 remains bound to the small RNA seems a huge waste of resources (Figure 14). Small RNAs titrate the RBP away from its original targets, i.e. translatable mRNAs. An explanation for this finding could be that binding of RBPs to small RNAs allows tighter control of organellar gene expression by the nucleus. When small RNAs do not release the RBP, consequently these RNA-bound PPRs cannot re-enter the pool of free PPRs. Under the assumption that organellar gene expression is limited by this free pool, expression and import into the organelle will more directly affect organellar gene expression and allow tighter nuclear control. PPR10 has a reported K_d in the sub-nanomolar range for its native target (Prikryl et al. 2011) and small RNAs indeed accumulate during leaf ageing (Sandra Gusewski, personal communication), supporting the idea that some PPR proteins are one times use only.

3.2 CP31A protects the *ndhF* mRNA against exonucleolytic decay

CP31A, a member of the chloroplast ribonucleoprotein (cpRNP) family, is essential for the accumulation of the *ndhF* mRNA in *Arabidopsis* (Tillich et al. 2009). As transcription rates were shown to be similar in *cp31a* and the WT, a defect in stability of the *ndhF* mRNA was proposed for *cp31a* (Tillich et al. 2009). The 3' ends for the *ndhF* mRNA were mapped in this thesis 470nt downstream of the *NdhF* stop codon (Figure 16, Figure 17). This 3' end was dependent on CP31A, while shorter and longer products were detectable in similar abundance to the WT in *cp31a* mutants (Figure 16). This transcript end is likely the dominant transcript 3' end in WT. It represents the most prominent band in the 3' RACE analysis, and the calculated length, including this long 3' UTR, fits the dominant signal in RNA gel blot analysis at 3.0knt (Figure 16, Figure 19). CP31A was shown to bind the *ndhF* mRNA *in vivo* by RNA-immunoprecipitation and chip analysis. Fine-mapping using oligonucleotide arrays revealed the highest enrichment close to the processed 3' end of *ndhF* (Kupsch et al. 2012). CP31A thus binds the *ndhF* mRNA close to the processing site affected in *cp31a* mutants. Identification of the exact binding site of CP31A in the *ndhF* mRNA could be achieved using *in vitro* binding assays or cross-linking and immunoprecipitation combined with sequencing of bound RNAs, iCLIP or PAR-CLIP approaches (Hafner et al. 2010, Konig et al. 2010). Additional evidence that the CP31A-mediated stabilization is conferred via sequences in the 3' UTR comes from the finding that antisense transcripts to *ycf1* that partially share the same sequence with the *ndhF* 3' UTR are similarly

reduced in *cp31a* mutants (Figure 19). The accumulation of 3' shortened RNA species for these antisense transcripts in *cp31a* mutants indicates that CP31A stabilizes *ycf1*as transcripts and likewise the *ndhF* mRNA against exonucleolytic decay from the 3' end (Figure 19). As discussed in the previous sections, a number of RBPs, mostly belonging to the class of PPR proteins, is implicated in RNA end processing and stabilization similar to CP31A. The transcript ends affected in mutants of these RBPs often overlap with small RNAs which represent their footprints. These footprints, where analyzed, are missing in complete knockouts of the respective RBP (Figure 9), (Hammani et al. 2012, Zhelyazkova et al. 2012a). The CP31A-dependent 3' end of *ndhF* and *ycf1* antisense transcripts overlaps with a small RNA (Figure 17). The detection of this small RNA by RNA gel blot or RNase protection is challenging as many small RNA isoforms with different 5' and 3' ends exist (Figure 18). Using an RNase protection assay, which is more sensitive than RNA gel blot analysis, small RNAs at the *ndhF* 3' end were shown to be strongly reduced in *cp31a* mutants and to lesser extent in *cp29a* mutants (Figure 18). Importantly, the small RNAs were not completely absent. This finding could indicate that either a paralogue of CP31A, CP31B, can at least in part complement the function of CP31A or that CP31A acts in a complex that is only partially destabilized when CP31A is missing. Best candidates for additional factors of this complex are PPR and PPR-like proteins. The PPR-SMR protein SVR7 could be part of this complex, as small RNAs at the 3' end of *ndhF* were found to be reduced and lack 3' extensions in a *svr7* mutant according to small RNA sequencing (Supplementary Figure 2). This hypothesis could be tested by analyzing *ndhF* and *ycf1* antisense transcripts in *svr7* mutants.

Potentially, CP31A does form complexes with PPR proteins also on other RNAs. Reduced stability of such complexes in *cp31a* mutants could explain the reduced editing efficiency seen at a number of editing sites (Tillich et al. 2009). Similarly under cold stress conditions CP31A could guide PPR proteins to targets, which show reduced stability in the cold when CP31A is not present (Kupsch et al. 2012).

3.3 Novel RNA-editing sites identified in *Arabidopsis*

3.3.1 Determination of editotypes by RNA-Seq

Massive parallel sequencing of cDNAs or short RNA-Seq allows the genome-wide investigation of a transcriptome and has been recently also applied to organellar transcriptomes (reviewed in Small et al. 2013). In this thesis, an RNA-Seq library was reanalyzed to elucidate the potential of RNA-Seq for the quantification and discovery of

RNA-editing events. The dataset contains about 30,000,000 raw reads per sample analyzed. As RNA-Seq datasets reflect the abundance of transcripts, read depth at the known 34 C→U editing sites differed substantially (Table 3), (Chateigner-Boutin and Small 2007). When the coverage was below 50 at a given editing-site the deviation between the two replicates investigated was in general higher. It is thus advisable to reach this minimum coverage for the least abundant RNA-editing site for a complete editotype. In the libraries investigated, the editing-site with the lowest abundance was *rpoC1*-21806. For a coverage of about 50 reads at this editing site an approximately 10fold higher sequencing depth would have been necessary. Thus starting with rRNA-depleted total RNA from *Arabidopsis* leaf tissue, about 300,000,000 raw reads per sample would be necessary for a complete, high confidence, chloroplast editotype. Even though sequencing costs are decreasing substantially, adjusting the coverage by sequencing RT-PCR products is advisable since more cost efficient (Bentolila et al. 2013). If detection of novel RNA-editing sites is intended, RNA-Seq offers a great opportunity for rapid detection. Especially when mitochondrial RNA-editing sites are in the focus of a study amplicon sequencing of RT-PCR products is advisable as mitochondrial transcripts were found less well covered in the dataset investigated (Ruwe et al. 2013). In general, both strand-specific high-throughput sequencing techniques using amplicon sequencing (Bentolila et al. 2013) and RNA-Seq (this thesis) report similar editing efficiency at plastid editing sites (Table 3). The highest deviation between the two datasets was found at the only editing site in the *clpP* gene, encoding a subunit of the plastid Clp protease. RNA-editing extend has been described to vary depending on developmental state and under stress conditions (Chateigner-Boutin and Hanson 2003, Karcher and Bock 1998). Plant material for the two studies was grown under slightly different conditions regarding day length and temperature and plant material was harvested at different developmental stages (Bentolila et al. 2013, Hotto et al. 2011). These differences can explain the slight deviations between the two datasets. In conclusion, RNA-Seq allows quantification of RNA editing in a strand-specific manner with the drawback of still relatively high costs due to sequencing of a majority of cDNAs without editing site and underrepresentation of lowly expressed transcripts. On the other hand RNA abundance, splicing status of intron containing genes and other RNA processing events can be measured in parallel with quantification of RNA editing (Hotto et al. 2015). Furthermore, RNA-Seq does not require a priori knowledge of RNA-editing sites.

3.3.2 Identification of promiscuous RNA-editing events

RNA-Seq has the potential to identify novel RNA-editing sites in expressed transcripts, which are not detectable using conventional techniques. A variant as abundant as 10% of a population is often not detectable by Sanger sequencing (Bentolila et al. 2013). Discrimination of RNA/DNA inconsistencies that result from RNA editing from those arising through technical artifacts is important. Polymeric tracts can result in polymerase slippage during library preparations and result in deletion or incorporation of additional nucleotides which can result in alignments pretending an RNA-editing event (Clarke et al. 2001, Falvey et al. 1976). Similarly, mapping artifacts resulting from low stringency applied, to allow mismatches due to RNA editing, can lead to miscalling. Post-transcriptionally added poly-A tails can be mapped to some A-T rich regions in the chloroplast genome. Similarly, nuclear transcripts encoded in NUPTs (nuclear plastid DNA), resulting from DNA transfer events from the chloroplast to the nuclear genome, can align (Michalovova et al. 2013). Mutations that occurred after nuclear transfer will then be detected as mismatches. All of the aforementioned cases were observed in mappings against the chloroplast genome. Every position that was detected to have a variant of at least 3% in both biological replicates was therefore checked manually. A large number of inconsistencies was identified in rRNAs and tRNAs. Both RNA species are known to be heavily modified in different domains of life (reviewed in Decatur and Fournier 2002, Novoa et al. 2012), but the knowledge about modifications in plastid tRNAs and rRNA is scarce. A number of potential modifying enzymes with plastid location have been described (Delannoy et al. 2009, Karcher and Bock 2009, Majeran et al. 2012, Tokuhisa et al. 1998). One tRNA modification an A→I modification in the wobble position was previously identified in *trnR*-ACG (Delannoy et al. 2009, Karcher and Bock 2009, Pfitzinger et al. 1990) and is also present in the dataset. It represents the only A→G mismatch in tRNA regions in the dataset investigated (Table 4), consistent with the prediction of only one adenosine to inosine RNA-editing event (Karcher and Bock 2009). In addition, in the *pnp* datasets two A→G mismatches have been identified in non-coding regions. One is situated in the intergenic region between *rbcL* and *accD* the other in the intron of *rpoC1*. Whether these indeed represent RNA-editing events like adenosine to inosine deamination needs further experimental support. The fact that these sites are only detected in the *pnp* mutants can be explained by the fact that free introns and extended 3' UTRs accumulate when the PNPase is absent (Castandet et al. 2013, Germain et al. 2011, Hotto et al. 2011).

Next to A→I editing, C→U deamination has been described for 34 sites in the chloroplast genome (Chateigner-Boutin and Small 2007). Additional ten such sites have been identified in RNA-Seq datasets in this thesis (Table 5). Six of these sites have been confirmed by sequencing individual cDNA clones (Table 5) and three by Cleaved Amplified Polymorphic Sequence (CAPS) analysis by Benoît Castandet (Ruwe et al. 2013). These ten sites arise very likely through RNA editing. These novel editing sites identified showed low C→U conversion rate (Table 5). Moreover, sites were identified predominantly in non-coding regions or were silent, i.e. do not change coding when edited. The question arises, whether editing is needed at these sites or is a result of promiscuous binding of RNA-editing factors, similar as described above for stabilizing PPR proteins like CRR2 (3.1.3).

3.3.3 Prediction of editing factors for promiscuous RNA-editing events

To provide evidence for the hypothesis that the newly identified RNA-editing sites represent off-targets of known RNA-editing factors, alignments of PPR repeats of trans-factors with potential *cis*-elements of the novel RNA-editing sites have been performed. 19 RNA-editing factors, belonging to the family of PLS-class PPR proteins, are required for editing at 30 of the 34 known RNA-editing sites in chloroplasts (Hammani et al. 2009, Wagoner et al. 2015, Yagi et al. 2013, Yap et al. 2015). Using a scoring matrix published recently (Yap et al. 2015), these 19 editing factors were aligned with the ten novel editing sites to predict possible binding. The *cis*-elements were aligned with the PPR repeats of the editing factors as described previously so that the terminal S motif aligns with the base at position -4 with regard to the edited C (Barkan et al. 2012, Takenaka et al. 2013a, Yagi et al. 2013). The result of this analysis is presented in Supplementary Table 1. Some of the scores for alignments of new *cis*-elements with RNA-editing factors are higher than for genetically determined target sites. Examples are the PPR/RNA pairs QED1 (OTP81)/ *ndhB*3'UTR-94622, CRR22/ *rps18* 3'UTR-68453 and CLB19/ *ycf3* Intron 2-43350. Testing one of these predicted pairs showed that CLB19 is indeed required for RNA editing at the novel editing site in the group II intron of *ycf3*, as measured by poisoned primer extension in *clb19* mutants. Additionally, recombinant CLB19 protein has a high affinity for the sequence upstream of the editing site (Dr. Peter Kindgren, personal communication). Whether editing or binding of CLB19 at this position in the intron has an influence on splicing needs to be determined. The novel RNA-editing site in the 3' UTR of *rps18* is special as the potential *cis*-element and the editing site is present in small RNA (C40 in

Supplementary Table 2). CRR22 can be predicted with high score to bind the sequence upstream of the edited C. Whether CRR22 is required for the editing reaction and/or for the accumulation of the small RNA that overlaps with the dominant 3' end of *rps18* as determined by 3' RACE (Ruwe and Schmitz-Linneweber 2012) needs further experimental evidence. Thus a number of known RNA-editing factors is predicted with high scores to bind sequences upstream of novel RNA-editing sites identified in this thesis. Whether the binding and/or editing at these sites is beneficial or just tolerated needs to be determined.

Finally it needs to be discussed, whether all previously known sites are required or whether some of the 34 sites in *Arabidopsis* result from promiscuous binding of PPR proteins. QED1 was shown to be required for five RNA-editing events in *Arabidopsis* chloroplasts (Wagoner et al. 2015). Two of these sites are only partially edited and are located in non-coding regions in the 3' UTR of *accD* and in intron I of *rps12*. Thus, these two sites share characteristic features with the ten novel RNA-editing sites. Even though RNA editing can be important for intron splicing as demonstrated in mitochondria, editing in intron I of *rps12* seems not to be required for efficient splicing (Castandet et al. 2010, Hammani et al. 2009). This indicates that the two sites could represent off-targets of QED1.

The question arises why some sites targeted by an RNA-editing factor show lower C→U conversion. The affinity of a PPR editing factor can influence RNA-editing efficiency (Kindgren et al. 2015). Additional factors might likewise be important. CP31A described above was shown to be required for efficient editing at several plastid sites (Tillich et al. 2009) and members of the RIP/MORF class of proteins have been shown to be required for efficient editing in plastids and mitochondria (Bentolila et al. 2012, Bentolila et al. 2013, Sun et al. 2013, Takenaka et al. 2012). Deep sequencing of RT-PCR products in RIP mutants showed that two novel editing sites present in the *ndhB* open reading frame, showed strongly increased RNA editing, while at previously known sites editing was decreased. This suggests that MORF/RIP proteins and possibly cpRNPs might help the editing machinery to distinguish between real and off-targets (Bentolila et al. 2013).

4 Material and Methods:

4.1 Materials

A list of suppliers can be found in the appendix (Supplementary Table 4).

4.1.1 Chemicals and Biochemicals

Chemicals and biochemicals were purchased from Carl Roth, Sigma-Aldrich and Thermo Scientific if not otherwise stated. All solutions were prepared with *A. bidest* (de-ionized, distilled water, PURELAB-Ultra-system, Veolia).

4.1.2 Plant material

Arabidopsis thaliana

Table 6: *Arabidopsis* lines used in this study

Line	Database entry	Mutant first described in
***crr2*-3**	SALK_030786	-
***crr2*-4**	SALK_046131	-
gun1-102	SAIL_290_D09	-
***hcf107*-2**	FLAG_DEI117	(Felder et al. 2001, Sane et al. 2005)
***hcf152*-1**	FLAG_CRM3	(Meierhoff et al. 2003)
***mrl1*-1**	SALK_072806	(Johnson et al. 2010)
***mrl1*-3**	SAIL_862_D12	-
***sot1*-2**	GK_840D06	-
***svr7*-2**	CSHL_GT20858	(Zoschke et al. 2013)
***svr7*-3**	SAIL_423_G09	(Zoschke et al. 2013)

The *gun1*-102, *sot1*-2, *svr7*-3 lines were obtained from Dr. Kate Howell and *mrl1*-3 from Dr. Sandra Tanz. The *mrl1*-1 line was obtained from Dr. Katia Wostrikoff. The *hcf107*-2 and *hcf152*-1 lines were obtained from Prof. Peter Westhoff. The lines *crr2*-3 and *crr2*-4 were ordered from NASC (Nottingham *Arabidopsis* Stock Centre) and genotyped by PCR analysis (4.2.3).

Zea mays

Co-immunoprecipitation of RNAs bound to PPR10 was performed from B73 maize.

4.1.3 Bacterial strains

For plasmid propagation of RT-PCR products from RACE experiments and for confirmation of novel C→U editing sites, the *E.coli* strains TOP10 (Life Technologies) and DH5α were used.

4.1.4 Oligonucleotides

DNA oligonucleotides were ordered as desalted or HPLC purified from Invitrogen, Sigma-Aldrich or Eurofins MWG Operon. RNA oligos for 5' and 3' RACE were synthesized by Illumina, Metabion and NEB. A list of oligonucleotide sequences can be found in Table 7.

Table 7: Oligonucleotides used in this thesis. T7 promoter sequences are underlined. P indicates 5' phosphate modification; idT indicates a 3'-3' linkage with deoxythimidine.

Primer name	Sequence	Comment
5' RACE		
Rumsh	GUGAUCCAACCGACGCGACAAGCUAAUGCAAGANNN(RNA)	Linker
5' SR Adaptor	GUUCAGAGUUCUACAGUCCGACGAUC(RNA)	Linker (*ndhA*, *rrn23*)
5AdapterRACE	GTTCAGAGTTCTACAGTCCGAC	*ndhA*, *rrn23* RACE
*ndhA*_5RACE	CCTGTTATGATTCCCAATACAAG	
23S precursor rev	CCTCGCCCTTAACTTTAAGGC	
Rumsh1	TGATCCAACCGACGCGAC	Adapter Primer
rps15 5′	CCAAATGTGAAGTAAGTCTTCG	
ndhB 5′	TATCCAGATAATAGGTAGGAGC	
psbC.T7	GTAATCGACTCACTATAGGGCCCCCAAAGGGAGATTTTAG	
3' RACE		
SRA 3′-Adapter	P-UCGUAUGCCGUCUUCUGCUUGUidT (RNA)	Linker for 3′RACE
AdapterRT primer	CAAGCAGAAGACGGCATA	RT
AdapterPCR primer	CAAGCAGAAGACGGCATACG	PCR
ycf1 3′RACE	AGCTTGTATGAATCGCTATTGG	
rps7 3′RACE	CGATGCCATACGCAAAAAGG	
ndhF 3'RACE	GTCGCATCTCTTCTTATCTGTTC	
*ycf1*as 3'RACE	CGAAAACGAGAGTTACAAATGG	
Confirmation of novel editing sites		
*ndhKJ*ed_rev	*tgatccaaccgacgcgac*NNNNGCTAGCCAAACGGACAAA	RT
*rps4*ed_rev	*tgatccaaccgacgcgac*NNNNGACCACAATGTATCAAATCC	RT
*ndhB*_ed_rev	*tgatccaaccgacgcgac*NNNNTCGTATACGTCAGGAGTC	RT
*atpH*_ed_rev	*tgatccaaccgacgcgac*NNNNAATTAGTCCTTCCCAAGG	RT
*ycf3*In_ed_rev	*tgatccaaccgacgcgac*NNNNGTTGTGTCGGTCCAAAAC	RT
Adapter Primer	*tgatccaaccgacgcgac*	PCR
*ycf3*In_ed_fwd	GTGCGACTATCTCCACTATAG	PCR
*ndhK-ndhJ*_ed_fwd	TAGACCTCAACAGGGTAATCG	PCR
*rps4*ed_for	GATAGGAAATGCGTCGGTTTG	PCR
*ndhB*ex1.rp	CCGATGGAGAGAAGAACCTATG	PCR

Primer name	Sequence	Comment
At-*atpH*_fw	ATGAATCCACTGGTTTCTGCTGC	PCR
Generation of templates for *in vitro* transcription		
*ycf2*as.T7	TAATACGACTCACTATAGGGATCCTCGTACATGGTG	*ycf2*as probe
ycf2 5'RACE	AATATCGATTGCTTGTTGAACC	*ycf2*as probe
matK.T7	TAATACGACTCACTATAGGGATCCTAATCTTAGGGAAAATGG	*matK* probe
matK.rp	GGCAACAGAGTTTTCTATATCCAC	*matK* probe
ndhB.T7	TAATACGACTCACTATAGGGTTGAATCGATCATCAGAAG	*ndhB-rps7* probe
*rps7*cRT1	GATCTCTTTCTCGAAACAAACG	*ndhB-rps7* probe
ycf2.T7	TAATACGACTCACTATAGGGAACAGATAGCAACAACAA	*ycf2* probe
ycf2.rp	GGATTAAGTGGAACGGAATTG	*ycf2* probe
*ndhF*3'UTR.T7	GTAATCGACTCACTATAGGGTGAGAAATTCTATGGCTCGAATC	*ndhF*3'UTR probe
*ndhF*3'UTR.rp	TCGAACGTGGAATTCATCATC	*ndhF*3'UTR probe
*ycf1*as.T7	GTAATCGACTCACTATAGGGAAGATGGAATCGACCAAACC	*ycf1*as probe
*ycf1*as.rp	GATTCTTCCCCGAGAGATTCC	*ycf1*as probe
*ndhF*short.T7	TAATACGACTCACTATAGGGAGAAGAGATGCGACTTCCAC	*ndhF* 3' probe
*ndhF*short.rp	TTTTTCACGCCGTCAATAAACC	*ndhF* 3' probe
Oligo probes for small RNA gel blots		
*rrn23*sRNAprobe	GAAAGATCTTATCAACGTCCATGAA	
*ndhA*sRNAprobe	GTATCGTCATAATATCAGCCAATTT	
*rpoA*assRNAprobe	GTCTACAATTGTCTCAAAAAATCCAATAT	
*rbcL*sRNAprobe	GCAATAAAACAAAACAACAAGGTCTACTCGACA	
psbH footprint	TTCATTACGATCTGTTGACTTTGTATACC	
psbH-petB footprint	CAGAAAAAAATTTCGCGGTCGAACTACC	
ndhB footprint	GTACATGCCAGATCATGAATTAGTAACT	
*matK*CDSprobe	GATTCTGTTCATACATTCGCAAAA	
*ycf2*_3probe	GTTCGCTGTTCAAGAATTCTTGTTT	
rps7 3sRNA	AGAGATCGATCAATTCCGATTTTTTCTTTTTCTAT	
Generation of templates for the RNase protection assay		
T7 with overlap	TAATACGACTCACTATAGG**GAGACAGG**	sequences in bold anneal
atpH footprint	TTGGTTGATTGTATCCTTAACCATTTCTTTTTTTTTGACA**CCTGTCTC**	
ndhF footprint	TAAAATGTGACCAATTAACCAACCAACAAAACTACTTG**CCTGTCTC**	
Sanger sequencing		
M13R	GGAAACAGCTATGACCATG	
pJet1.2rev	AAGAACATCGATTTTCCATGGCAG	
Oligonucleotides used as size markers		
ndhB_ed_7_rev_in	ATGCAGTATCGTCCTAGTCAGGGTAGGAATTTCTCAAACGAACC	44mer DNA
SF_C2 fSal2	GGACTGTCGACCATTATGGGGAAACCCTTTACG	33mer DNA
ycf1as.rp_Nt	GGTAGAAATCCACTGATTGTCC	22mer DNA
Rumsh	GUGAUCCAACCGACGCGACAAGCUAAUGCAAGANNN	36mer RNA
SRA 5′-Adapter	GUUCAGAGUUCUACAGUCCGACGAUC	26mer RNA
Genotyping of T-DNA insertion lines		
*crr2*rev	TCGAATTTGAGGGCACAATGAA	
*crr2*fwd	AATGCATGACCGGGATGTTG	
LBa1	TGGTTCACGTAGTGGGCCATCG	

4.1.5 Antibodies

The affinity-purified anti-PPR10 antibody (polyclonal) was obtained from Prof. Alice Barkan (Pfalz et al. 2009). The anti-PPR4 antibody is an affinity-purified polyclonal antibody (Schmitz-Linneweber et al. 2006).

4.2 Methods

4.2.1 Sterilization of solutions and inactivation of GMOs

Sterilization of solutions and inactivation of genetically modified organisms was performed by autoclaving for 20min at 120°C/ 55kPa using a Varioklav 75 S steam autoclave (Thermo Scientific).

4.2.2 Plant growth conditions

Plants grown on soil

Arabidopsis and maize was grown on a soil (Einheitserde GS90; Gebrüder Patzer) and vermiculite mixture (4:1; 2-6mm, Floragard). Maize was grown at 28°C, 16h light/8h dark cycle, ~120µmol × m^{-2} × s^{-1}. *Arabidopsis* was grown at 23°C at long day conditions (16h light/8h dark) at light intensities of ~120µmol × m^{-2} × s^{-1}.

Plants grown on MS-medium containing sugar

The *hcf107*-2 and *hcf152*-1 lines were grown on MS-medium containing 3% (w/v) sucrose. Heterozygous seeds were surface sterilized in sterilization solution for 7min and washed five times in autoclaved water. Plants were grown at ~60µmol × m^{-2} × s^{-1} at 23°C under long-day conditions. Homozygous plants were identified by their high chlorophyll fluorescence phenotype under UV-light.

Sterilization solution:	32% (v/v) DanKlorix (Colgate-Palmolive), 0.8% N-lauryl-sarcosine.
MS-medium:	0.44 % (w/v) Murashige and Skoog Media (Duchefa), 0.05% (w/v) MES, 0.5 % (w/v) plant agar (Duchefa), 3% (w/v) sucrose; pH 5.7 with KOH (Murashige and Skoog 1962).

4.2.3 Genotyping

For genotyping of T-DNA insertion lines (4.1.2), DNA was isolated by homogenizing 5-10mg leaf tissue in a microfuge tube using a pestle, following a slightly modified protocol (Edwards et al. 1991). The tissue was lysed in 700µl DNA extraction buffer. Insoluble material was removed by centrifugation and nucleic acids precipitated by addition of one volume isopropyl alcohol. After precipitation by centrifugation, the pellets were washed with 70% ethanol. DNA was resuspended in *A. bidest*.

Two PCR reactions (4.2.6) were used to analyze zygosity, one PCR with primers spanning the proposed insertion detecting the wild-type (WT) allele and one with a gene-specific primer and one primer located in the T-DNA left border. PCR products were separated on agarose gels (4.2.7).

DNA extraction buffer: 200mM Tris-HCl pH 7.5, 250mM NaCl, 25mM EDTA, 0.5% SDS

4.2.4 RNA Isolation

Standard isolation of total RNA was performed with the TRIzol reagent (Life Technologies) following the manufactures instructions after homogenization with either a ball mill (Mixer Mill 400, Retsch) or mortar and pestle. RNA was stored in *A. bidest* at -80°C. For the analysis of RNA accumulations in *crr2* and *sot1* mutants (2.1.4.1 and 2.1.4.2) RNA was isolated using a column-based protocol. Plant material was flash frozen in liquid nitrogen and homogenized using a ball mill (Mixer Mill 400, Retsch). Plant material was lysed in 1ml lysis solution per 100mg plant tissue. RNA was isolated following the manufactures instructions for the Direct-zol™ RNA MiniPrep Kit (Zymo Research). RNA isolation for small RNA sequencing and 5' RACE analysis in *sot1* mutants for *ndhA* and *rrn23* (4.2.11, 4.2.18) was performed using the miRNeasy Mini Kit (QIAGEN) following the manufactures instructions. An optional homogenization step using spin columns (QIAshredder, QIAGEN) was included.

Lysis solution: 48% (v/v) water-saturated phenol, 2M guanidinium thiocyanate, 25mM Tris-HCl pH 4.5, 5mM EDTA, 0.12% N-lauryl-sarcosine, 2.12% (v/v) isoamyl alcohol, 0.1% (w/v) hydroxyquinoline, 0.5% (v/v) β-mercaptoethanol.

4.2.5 Spectroscopic measurement of nucleic acid

Quantity and purity of nucleic acids in solution was determined using a UV spectrophotometer (NanoDrop 1000, PEQLAB). RNA integrity was judged from integrity of rRNA bands in agarose gels (4.2.7).

4.2.6 Polymerase chain reaction (PCR)

Recombinant DNA polymerase I from *Thermus aquaticus* was purified from *E.coli* strain DH5α using a published protocol (Desai and Pfaffle 1995). A standard PCR reaction contained a 1X PCR buffer, 0.2mM dNTPs (Thermo Scientific), 0.2µM forward and reverse primer, *Taq* Polymerase 1:50 dilution and cDNA or DNA template (1:50 dilution). A temperature profile for a standard PCR reaction is shown below. Annealing temperatures were determined using the online tool "NEB Tm calculator" (http://tmcalculator.neb.com). Denaturation, annealing and elongation were repeated for 25-35 cycles.

	Temperature	Time
Initial denaturation	94°C	3min
Denaturation	94°C	30sec
Annealing	45-58°C	30sec
Elongation	72°C	1min/kb
Final elongation	72°C	5min

10X PCR buffer: 200mM Tris-HCl pH 8.8, 100mM KCl, 100mM $(NH_4)_2SO_4$, 20mM $MgSO_4$, 1% Triton X-100

4.2.7 Agarose gel electrophoresis

For separation of nucleic acids on native agarose gels, 1-3% agarose (Biozym) was melted in 1X TAE buffer in a microwave oven. Ethidium bromide (final concentration: 0.2µg/ml) was added and the agarose was allowed to gel at room temperature. DNA samples were mixed with one volume 10X sample buffer per nine volumes sample. RNA samples were mixed with at least one volume RNA sample buffer (4.2.13) and heated for 5min at 75°C. Samples and dsDNA Markers (GeneRuler 1kb DNA Ladder, GeneRuler 100bp Plus DNA Ladder, Thermo Scientific) were run at 5-10V/cm in 1X TAE as running buffer. Gels were documented under UV light (302nm; Gel Doc XR™, Bio-Rad).

1X TAE buffer: 40mM Tris, 20mM acetic acid, 1mM EDTA

10X Sample buffer: 0.42% bromophenol blue, 0.42% xylene cyanol, 25% ficoll (Type 400)

4.2.8 cDNA synthesis for confirmation of novel editing sites

For the confirmation of novel RNA-editing sites (2.3.2.2), RNA isolated from three week old WT plants was treated with 10 units DNase I (Roche) followed by standard phenol-chloroform extraction and ethanol precipitation (Sambrook and Russell 2001). RNA was reverse transcribed using SuperScript III reverse transcriptase (Life Technologies) according to the manufacturer's manual, with gene-specific primers (Table 7) that contain four random nucleotides to distinguish individual reverse transcription events. This barcode was preceded by a binding site for a primer for PCR amplification (4.2.6).

4.2.9 Transformation of chemical competent *E.coli*

50µl chemically competent *E.coli* (4.1.3) cells were thawed on ice and incubated with a maximum of 5µl ligation reaction for 30min on ice. A heat shock at 42°C was carried out for 30 seconds in a water bath. Cells were allowed to recover in SOC medium at 37°C for 30 to 60min before plating on LB agar plates containing 100µg/ml carbenicillin.

SOC medium: 2% (w/v) tryptone, 0.5% (w/v) yeast extract, 10mM NaCl, 2.5mM KCl, 10mM $MgSO_4$, 10mM $MgCl_2$, 20mM glucose, pH 7.0 with NaOH

LB agar plates: 1% (w/v) tryptone, 0.5% (w/v) yeast extract, 1% (w/v) NaCl, 1.5% bacto agar, pH 7.0 with NaOH

4.2.10 Preparation of plasmids from *E.coli*

Single *E.coli* colonies were grown overnight at 37°C in LB medium containing 100µg/ml carbenicillin. Plasmids were purified using the GeneJET Plasmid Miniprep Kit (Thermo Scientific).

LB medium: 1% (w/v) tryptone, 0.5% (w/v) yeast extract, 1% (w/v) NaCl, pH 7.0 with NaOH

4.2.11 5' and 3' RACE

For determination of transcript ends in chloroplasts, a rapid amplification of cDNA ends (RACE) approach was conducted. Total RNA was ligated to small RNA or DNA oligonucleotides (Table 7) with T4 RNA Ligase I (NEB), according to the manufacturer's instructions. For 5' RACE of *ndhA* and precursors of *rrn23* one sample was treated with tobacco acid pyrophosphatase (TAP, Epicenter) to convert 5' triphosphorylated primary transcript ends into monophosphate ends to allow ligation. A second sample was untreated to distinguish between primary and secondary ends. After TAP treatment and after linker ligation, RNA was purified with standard phenol-chloroform extraction and ethanol precipitation with 0.3M sodium acetate (Sambrook and Russell 2001). RNA was reverse transcribed into cDNA with SuperScript III reverse transcriptase (Life Technologies) using random primers for 5' RACE and an adapter-specific primer for 3' RACE (Table 7).

PCR amplification of ligation products was performed as described in 4.2.6. PCR products were eluted from agarose gels using the GeneJET Gel Extraction Kit (Thermo Scientific) or the QIAquick Gel Extraction Kit (QIAGEN) according to the manufacturer's manuals. PCR products were cloned with the CloneJET PCR Cloning Kit (Thermo Scientific) or the pGEM-T Easy vector system (Promega) and transformed in *E.coli* cells (4.2.9). Single colonies were screened for correct insert size by PCR. PCR products were purified and sequenced with primer M13R for the pGEM-T Easy vector by Macrogen (*ndhA* and *rrn23* 5' RACE, 2.1.4.1). Clones containing the pJet1.2 vector were propagated and plasmids purified as described in 4.2.10. Plasmids were Sanger sequenced with primer pJet1.2rev by SMB.

4.2.12 RNA gel blot analysis using agarose gels

RNA agarose gel electrophoresis

For analysis of long RNAs, i.e. mRNAs and long non-coding RNAs by RNA gel blot analysis, total RNA was separated on agarose gels containing formaldehyde as a denaturing agent. Concentrations of agarose varied between 1-1.3%. RNA was denatured in at least 2.5 volumes RNA sample buffer for 15min at 70°C. An RNA ladder served as a size marker and was treated the same way (RiboRuler High Range, Thermo Scientific). Samples and ladder were separated in an ice-cooled horizontal agarose gel-electrophoresis system with buffer circulation using 1X MOPS buffer as running buffer (for some gels the running

buffer was supplemented with ~1.85 % formaldehyde). The voltage was set constant at 5-7V/cm.

RNA agarose Gel: 1.2-1.56g agarose (Certified™ molecular biology agarose, Bio-Rad) in 88ml H_2O, 12ml 10X MOPS (pH 7.0), 20ml formaldehyde solution (37%)

RNA sample buffer: 65% (v/v) deionized formamide, 22% formaldehyde solution (37%), 13% (v/v) 10X MOPS buffer, trace amounts of bromphenol blue and xylene cyanol, optional: ethidium bromide (0.05µg/µl)

10X MOPS buffer: 200mM MOPS, 10mM EDTA, 80mM NaOAc, pH 7.0 with NaOH

Capillary transfer of RNA to nylon membranes

RNA separated in denaturing agarose gels was blotted to nylon membranes (Hybond-N, GE Healthcare) by passive transfer with 5XSSC (Sambrook and Russell 2001). RNA was fixed on membranes by UV radiation (250mJ/cm^2; GS Gene Linker, Bio-Rad). To control transfer and equal loading, membranes were stained with methylene blue solution for ~2min and destained in water.

5X SSC: 0.75M NaCl, 0.075M sodium citrate , pH 7.0

Methylene blue solution: 0.3M NaOAc (pH 5.2), 0.03% (w/v) methylene blue

Preparation of ^{32}P-labeled RNA probes

Templates for *in vitro* transcription were amplified by PCR (4.2.6), using a reverse primer that introduces a T7 promoter sequence (Table 7). PCR products were purified using the GeneJET PCR Purification Kit (Thermo Scientific). An *in vitro* transcription reaction was set up according to the manufacturer's instructions (T7 RNA polymerase, Thermo Scientific). 50µCi α-^{32}P-UTP (Hartmann Analytics) were used to label the RNA probe. Unincorporated nucleotides were removed using gel filtration columns (illustra MicroSpin G-50, GE Healthcare).

Hybridization, stringency washes and signal detection

Membranes were prehybridized in Church buffer at 68°C for at least 1h. After prehybridization, radiolabeled probes were added and membranes hybridized overnight at 68°C. Stringency washes were performed at 68°C by reducing salt concentrations in consecutive washes. Membranes were washed for 20min in 0.5X SSC, 0.1% SDS followed by 0.2X SSC, 0.1% SDS and 0.1X SSC, 0.1% SDS. Signals were detected using a phosphoimaging system (PMI FX, Imaging Screen-K, Quantity-One-Software, Bio-Rad).

Church buffer: 0.5M sodium phosphate buffer (pH 7.0), 7% (w/v) SDS, 1mM EDTA

4.2.13 RNA gel blot analysis of small RNAs

RNA polyacrylamide gel electrophoresis

For the detection of small RNAs by RNA gel blot, purification of probes for RNase protection assays and for the size selection of small RNAs for small RNA sequencing, RNA was separated by size in denaturing polyacrylamide gels. Urea served as a denaturing agent. RNA was denatured in at least 1 volume of RNA sample buffer for 10min at 75°C. Gels were prerun at 25-30V/cm in 1X TBE (Mini-PROTEAN® system, Bio-Rad). DNA Oligonucleotides were used as size markers, treated in parallel with RNA samples, taking into account that DNA migrates about 10% faster than RNA of the same size (Sambrook and Russell 2001). Samples were run at 25-30V/cm until the bromphenol blue, present in the sample buffer, reached the bottom of the gel. Gels were stained in an ethidium bromide solution (~0.2µg/ml in 0.5X TBE) for 1-2min and briefly rinsed in 0.5X TBE before documentation under UV light (302nm; Gel Doc XR™, Bio-Rad).

RNA sample buffer: 95% formamide, 1mM EDTA, 0.02% SDS, traces of bromphenol blue and xylene cyanol

10X TBE: 0.89M Tris, 0.89M boronic acid, 20mM EDTA

RNA gel: 1X TBE, 12-15% acrylamide (29:1 acrylamide:bis-acrylamide), 8M urea (MP Biomedicals), 0.5% (v/v) TEMED, 0.05% (w/v) APS

RNA transfer and chemical cross-linking

RNA was transferred to nylon membranes (Hybond-N, GE Healthcare) in 0.5X TBE using the Mini-PROTEAN® electrophoresis system (Bio-Rad) for 1h at 80V. RNA was chemically cross-linked to the membrane using an 1-ethyl-3-(3-dimethylaminopropyl) carbodiimide (EDC) cross-linking reagent (Pall and Hamilton 2008). In the cross-linking reaction, 5' phosphorylated oligonucleotides are covalently coupled to amine groups on the nylon membrane. Chemical cross-linking increases the sensitivity of RNA gel blots for small RNAs (<40nt) by a factor of up to 50 (Pall et al. 2007). Membranes were briefly washed in *A. bidest.* and RNA stained with methylene blue to control for efficient transfer (4.2.12).

EDC cross-linking reagent: 0.16M 1-ethyl-3-(3-dimethylaminopropyl) carbodiimide in 0.13M 1-methylimidazole, pH 8.0 with HCl

Preparation of ^{32}P-endlabeled DNA oligonucleotides

DNA oligonucleotides (30-50pmol) used as probes in small RNA gel blot analysis were end labeled with polynucleotide kinase (PNK, Thermo Scientific) according to the manufacturer's instruction with 50µCi γ-^{32}P-ATP (Hartmann Analytics). Nucleotides were removed using gel filtration columns (illustra MicroSpin G-25, GE Healthcare). DNA oligos were denatured at 95°C for 5min and directly transferred to ice.

Hybridization and washing conditions

EDC cross-linked membranes were prehybridized at 37°C for at least 1h in Church buffer (4.2.12). Oligonucleotide probes were added and hybridization was allowed to occur overnight. Membranes were washed twice in 1XSSC, 0.1% SDS at 37°C for 10min. Signals were detected using a phosphoimaging system (PMI FX, Imaging Screen-K, Quantity-One-Software, Bio-Rad).

4.2.14 RNase protection assay

Preparation of radiolabeled probes

Templates for radioactive *in vitro* transcription were synthesized by hybridization of two DNA oligonucleotides (200pmol) in 1X annealing buffer in a 10µl reaction by heating for 5min at 70°C followed by incubation at room temperature for 5min. The overlap

between the two oligos consisted of eight consecutive bases. Annealed oligos were filled-up by Klenow Fragment, exo- (Thermo Scientific) by adding 2µl 10XTango buffer (Thermo Scientific), 0.5µl dNTPs (2mM each), 2.5U Klenow Fragment, exo- and water to 20µl at 37°C for 30min. 2µl of these fill-in reactions served as templates for radioactive *in vitro* transcriptions with α-^{32}P-UTP (Hartmann Analytic) and T7 RNA Polymerase (Thermo Scientific) according to the manufacturers manual with the exception that no unlabeled UTP was used. Templates were digested by addition of 2U Turbo DNase (Life Technologies) at 37°C for 15min. Probes were gel-purified on 12% Urea polyacrylamide gels (4.2.13). RNA was eluted from the gel slice containing the full-length probe with 125µl probe elution buffer (mirVana™ miRNA Detection Kit, Life Technologies).

10X Annealing buffer: 1M NaCl, 100mM Tris-HCl pH 7.5

Hybridization and RNase digestion

RNase protection assays were performed using the mirVana™ miRNA Detection Kit (Life Technologies) essentially as described in the manual. To facilitate precipitation of protected fragments, 5µg yeast RNA was added to RNase digested samples during precipitation. Precipitated RNAs were separated in 12% denaturing polyacrylamide gels (4.2.13) alongside end-labeled RNA oligonucleotides or a single-stranded DNA ladder (Low Molecular Weight Marker, Affimetrix). Gels were dried on a Model 583 Gel Dryer (Bio-Rad) and signals were detected using a phosphoimaging system (PMI FX, Imaging Screen-K, Quantity-One-Software, Bio-Rad).

4.2.15 Isolation of stroma fraction from intact chloroplasts

Intact chloroplasts from 10 day old maize seedlings were isolated as previously described (Voelker and Barkan 1995). Intact chloroplasts were lysed in small amounts of extraction buffer (200-400µl) by forcing chloroplasts through a 24 gauge needle about 30 times. Stroma was separated from membranes by centrifugation at 40.000×g for 30min. Protein concentration was measured using the Bio-Rad protein assay (Bio-Rad). Stroma fractions were stored in 10% Glycerol at -80°C.

Extraction buffer: 2mM DTT, 200mM KOAc, 30mM HEPES-KOH, pH 8.0, 10mM MgOAc, 1X Protease Inhibitor Cocktail, EDTA-free (Roche)

4.2.16 RNA co-immunoprecipitation and RNA isolation

200-500µg stromal protein fractions were diluted with co-immunoprecipitation buffer in a 1:1 ratio. This solution was incubated at 4°C for 1h with 5µl of affinity purified antibody against PPR10 or PPR4 (4.1.5). Antibodies were captured with 50µl Dynabeads Protein G (Life Technologies) prewashed in co-immunoprecipitation buffer. Beads were washed three times in 500µl co-immunoprecipitation buffer. Supernatants and pellets in co-immunoprecipitation buffer were supplemented with SDS and EDTA to reach a final concentration of 1% SDS and 5mM EDTA. RNA was isolated from supernatant and pellet fractions using standard phenol-chloroform isolation and ethanol precipitation (Sambrook and Russell 2001).

Co-immunoprecipitation buffer: 150mM NaCl, 20mM Tris-HCl pH 7.5, 1mM EDTA, 5mM $MgCl_2$, 0.5% Nonidet P-40, 5µg/ml aprotinin

4.2.17 Preparation of libraries for small RNA sequencing

For sequencing of small RNAs in mutants of RBPs (2.1.4), 10µg total RNA was size separated in 12% urea polyacrylamide gels (4.2.13) alongside two single-stranded RNA markers (microRNA Marker, Low Range ssRNA Ladder, NEB). Gels were stained in SYBR® Gold Nucleic Acid Gel Stain (Life Technologies) diluted in 1XTBE. The gels were cut between the 15 and 50nt marker bands and RNA eluted in 0.3M NaCl overnight. 15µg GlycoBlue™ Coprecipitant (Life Technologies) was added and RNA precipitated by addition of 2.5 volumes 96% ethanol. Pellets after precipitation by centrifugation were washed in 80% ethanol and air-dried. RNA was resuspended in *A.bidest.* and libraries were prepared according to the manual for the NEBNext Multiplex Small RNA Library Prep Set for Illumina (NEB) with 12 cycles of PCR amplification. PCR products were purified using the QIAquick® PCR Purification Kit (QIAGEN). Libraries were inspected on a 2100 Bioanalyzer (Agilent) using a DNA 1000 chip. Individual libraries were quantified on a Qubit™ Fluorometer (Life Technologies) using the Qubit® dsDNA HS Assay Kit (Life Technologies). Same amounts of individual libraries were pooled and purified in a native 5% Mini-PROTEAN® TBE Gel (Bio-Rad) according to the manual for the NEBNext Multiplex Small RNA Library Prep Set for Illumina (NEB). The pooled libraries were quantified by qPCR using the KAPA SYBR® FAST LightCycler 480 qPCR Kit (Kapa Biosystems) on a LightCycler® 480 System (Roche).

4.2.18 Small RNA sequencing

Small RNA libraries were sequenced on a MiSeq Desktop Sequencer (Illumina) using the MiSeq Reagent Kits v3 (150 cycles, Illumina) according to the manufacturers instruction. For higher read count an additional run on a HiSeq 1500 (Illumina) was carried out. The individual libraries were adjusted according to the reads obtained from the MiSeq run and purified and quantified as described in 4.2.17. Quantification by qPCR and sequencing was carried out by Dr. Kate Howell.

4.2.19 Bioinformatic analysis of small RNA sequencing data

Adapter trimming

Adapter sequences which are found at the 3' end of cloned small RNAs were trimmed with the cutadapt tool (Martin 2011) with following parameters:

```
-a "adapter sequence" -q 15 -m 15
```

This removes first low quality bases below a Phred score of 15 and then searches for the adapter sequence specified at the 3' end of reads. A minimum of 3 nucleotides at the 3' end of the read need to align with the first bases of the adapter to be trimmed.

Mapping

cDNA sequences were mapped against the *Arabidopsis* nuclear and organellar genomes (TAIR10 release) available from TAIR website (Lamesch et al. 2012). The short read mapper bowtie (Langmead et al. 2009) was used with the following parameters:

```
-a --best --strata -v 2 --sam
```

These settings let bowtie report all (`-a`) best (`--best --strata`) alignments possible which have a maximum of two mismatches (`-v 2`). The output format is a sam file. Sam files were converted to bam files and subsequently sorted and indexed using SAMtools (Li et al. 2009a). Coverage graphs were extracted from mappings using BEDTools (Quinlan and Hall 2010). For extraction of overall coverage graphs parameters were

```
genomecov -strand + -ibam inputfile.bam -bg > output.bdg
```

for forward strand and

```
genomecov -strand - -ibam inputfile.bam -bg > output.bdg
```

for the reverse strand. Extraction of only the 5' positions of all alignments were performed with the additional parameter `-5` and `-3` for 3' ends of alignments respectively. To

normalize, reads per million mapped reads (against the chloroplast genome) were extracted by scaling using the option `-scale`.

Small RNA extraction

For the extraction of small RNAs from small RNA mappings a pipeline was developed together with Gongwei Wang, who implemented the pipeline in R making use of the Bioconductor infrastructure (Lawrence et al. 2013). The pipeline is part of his PhD work and will be described in his thesis. In brief, 5' and 3' positions of small RNAs mapping to a chromosome are extracted from a sorted and indexed BAM file. The maximum in a window of 15nt is recorded for 5' and 3' ends separately. The first filter criterion is on read number and should be adjusted to the sequencing depth (40 for chloroplast and 60 for mitochondria using the dataset described in section 2.1.1). The second filter is on sharpness of the ends. For this the counts are divided by the coverage and only ends with values above 0.5 are retained. Thus the local background is considered which varies dramatically in the genome. As a last criterion the shape of a small RNA is used. Less reads are expected to have alignment ends in the region of the small RNA. Thus the number of alignment ends found within 15nt inside the small RNA had to be below 20% of the identified end plus the count for the two neighboring nucleotides for chloroplasts and below 50% for mitochondria. Visually spoken, this last criterion allows only peaks with relatively flat tops when looking at small RNA coverage (see for example Figure 8). Finally, as the aforementioned algorithms detect only one end of a small RNA the second end is determined by looking in a window of 15-50nt up- or downstream, dependent on the type of end, for the most dominant end of the other typ. In other words, if a sharp 3' was detected, the most dominant 5' end is identified in a window of 15-50nt upstream. This additional end does not need to fulfill the criteria above, but many small RNAs are identified by both a sharp 5' and 3' end.

Comparison of mutant and WT small RNA mappings

To extract differences in small RNA mappings, a constant factor of 0.1 was added to normalized counts from 5' and 3' ends. This translates into approximately one alignment end added at each genome position. This removes the problem of dividing by zero. WT values were divided by mutant values at each genome position and values above 20 were reported using the Integrated Genome Browser (Nicol et al. 2009).

4.2.20 Quantification of RNA editing by RNA-Seq

For the quantification of RNA editing and identification of potential new editing sites, RNA-Seq datasets from WT and *pnp* mutant tissue (Hotto et al. 2011) were reanalyzed using the CLC Genomics Workbench (Version 5.1).

Quality and adapter trimming

Low quality bases were removed using the default parameters allowing a maximum of one ambiguity. When adapter sequences were present they were removed with following parameters:

```
mismatch cost: 3, gap cost: 2, minimum score: 15, minimum end score: 2
```

Mapping

Trimmed reads were mapped in a strand-specific manner to the chloroplast genome (NCBI: NC_000932). The positions of known editing sites were manually converted from C to Y in the reference sequence to allow equal mapping of edited and unedited transcripts. Mapping parameters were:

```
Minimum length fraction: 90%, minimum similarity fraction: 80%
```

Quantification of RNA editing and identification of DNA-RNA conflicts

SNP detection was performed to extract DNA-RNA conflicts which include the known RNA editing sites. SNPs were called when the frequency of a non-DNA-encoded base exceeded 3% and the coverage exceeded 10 reads. The average Phred score at the position of the SNP and ten neighboring bases had to be above 20. This analysis resulted in separate tables for the two replicates of the WT and *pnp* mutants. Only SNPs present in both replicates were further considered. All SNPs were manually curated for potential PCR artifacts occurring in homopolymeric stretches (Clarke et al. 2001) or mapping artifacts resulting from mappings of nuclear or mitochondrial-encoded sequences. A list containing all identified sites is available as supplementary dataset 3 in Ruwe et al. (2013).

References

Adachi, Y., Kuroda, H., Yukawa, Y. and Sugiura, M. (2012) Translation of partially overlapping psbD-psbC mRNAs in chloroplasts: the role of 5'-processing and translational coupling. *Nucleic acids research*, **40**, 3152-3158.

Allison, L.A., Simon, L.D. and Maliga, P. (1996) Deletion of rpoB reveals a second distinct transcription system in plastids of higher plants. *The EMBO journal*, **15**, 2802-2809.

Arikit, S., Zhai, J. and Meyers, B.C. (2013) Biogenesis and function of rice small RNAs from non-coding RNA precursors. *Current opinion in plant biology*, **16**, 170-179.

Babiychuk, E., Vandepoele, K., Wissing, J., Garcia-Diaz, M., De Rycke, R., Akbari, H., . . . Kushnir, S. (2011) Plastid gene expression and plant development require a plastidic protein of the mitochondrial transcription termination factor family. *Proceedings of the National Academy of Sciences of the United States of America*, **108**, 6674-6679.

Backert, S., Lynn Nielsen, B. and Börner, T. (1997) The mystery of the rings: structure and replication of mitochondrial genomes from higher plants. *Trends in plant science*, **2**, 477-483.

Barkan, A. (1989) Tissue-dependent plastid RNA splicing in maize: transcripts from four plastid genes are predominantly unspliced in leaf meristems and roots. *The Plant cell*, **1**, 437-445.

Barkan, A. (2011) Expression of plastid genes: organelle-specific elaborations on a prokaryotic scaffold. *Plant physiology*, **155**, 1520-1532.

Barkan, A., Rojas, M., Fujii, S., Yap, A., Chong, Y.S., Bond, C.S. and Small, I. (2012) A combinatorial amino acid code for RNA recognition by pentatricopeptide repeat proteins. *PLoS genetics*, **8**, e1002910.

Barkan, A. and Small, I. (2014) Pentatricopeptide repeat proteins in plants. *Annual review of plant biology*, **65**, 415-442.

Barkan, A., Walker, M., Nolasco, M. and Johnson, D. (1994) A nuclear mutation in maize blocks the processing and translation of several chloroplast mRNAs and provides evidence for the differential translation of alternative mRNA forms. *The EMBO journal*, **13**, 3170-3181.

Bendich, A.J. (2004) Circular chloroplast chromosomes: the grand illusion. *The Plant cell*, **16**, 1661-1666.

Bentolila, S., Heller, W.P., Sun, T., Babina, A.M., Friso, G., van Wijk, K.J. and Hanson, M.R. (2012) RIP1, a member of an Arabidopsis protein family, interacts with the protein RARE1 and broadly affects RNA editing. *Proceedings of the National Academy of Sciences of the United States of America*, **109**, E1453-1461.

Bentolila, S., Oh, J., Hanson, M.R. and Bukowski, R. (2013) Comprehensive high-resolution analysis of the role of an Arabidopsis gene family in RNA editing. *PLoS genetics*, **9**, e1003584.

Binder, S., Stoll, K. and Stoll, B. (2013) P-class pentatricopeptide repeat proteins are required for efficient 5' end formation of plant mitochondrial transcripts. *RNA biology*, **10**, 1511-1519.

Bobrovskyy, M. and Vanderpool, C.K. (2013) Regulation of bacterial metabolism by small RNAs using diverse mechanisms. *Annual review of genetics*, **47**, 209-232.

Bollenbach, T.J., Lange, H., Gutierrez, R., Erhardt, M., Stern, D.B. and Gagliardi, D. (2005) RNR1, a 3'-5' exoribonuclease belonging to the RNR superfamily, catalyzes 3' maturation of chloroplast ribosomal RNAs in Arabidopsis thaliana. *Nucleic acids research*, **33**, 2751-2763.

Boulouis, A., Raynaud, C., Bujaldon, S., Aznar, A., Wollman, F.A. and Choquet, Y. (2011) The nucleus-encoded trans-acting factor MCA1 plays a critical role in the regulation of cytochrome f synthesis in Chlamydomonas chloroplasts. *The Plant cell*, **23**, 333-349.

Boussardon, C., Salone, V., Avon, A., Berthome, R., Hammani, K., Okuda, K., . . . Lurin, C. (2012) Two interacting proteins are necessary for the editing of the NdhD-1 site in Arabidopsis plastids. *The Plant cell*, **24**, 3684-3694.

Cai, W., Okuda, K., Peng, L. and Shikanai, T. (2011) PROTON GRADIENT REGULATION 3 recognizes multiple targets with limited similarity and mediates translation and RNA stabilization in plastids. *The Plant journal : for cell and molecular biology*, **67**, 318-327.

Castandet, B., Choury, D., Begu, D., Jordana, X. and Araya, A. (2010) Intron RNA editing is essential for splicing in plant mitochondria. *Nucleic acids research*, **38**, 7112-7121.

Castandet, B., Hotto, A.M., Fei, Z. and Stern, D.B. (2013) Strand-specific RNA sequencing uncovers chloroplast ribonuclease functions. *FEBS letters*, **587**, 3096-3101.

Chateigner-Boutin, A.L. and Small, I. (2007) A rapid high-throughput method for the detection and quantification of RNA editing based on high-resolution melting of amplicons. *Nucleic acids research*, **35**, e114.

Chateigner-Boutin, A.L. and Hanson, M.R. (2003) Developmental co-variation of RNA editing extent of plastid editing sites exhibiting similar cis-elements. *Nucleic acids research*, **31**, 2586-2594.

Clarke, L.A., Rebelo, C.S., Goncalves, J., Boavida, M.G. and Jordan, P. (2001) PCR amplification introduces errors into mononucleotide and dinucleotide repeat sequences. *Molecular pathology : MP*, **54**, 351-353.

Colcombet, J., Lopez-Obando, M., Heurtevin, L., Bernard, C., Martin, K., Berthome, R. and Lurin, C. (2013) Systematic study of subcellular localization of Arabidopsis PPR proteins confirms a massive targeting to organelles. *RNA biology*, **10**, 1557-1575.

Coquille, S., Filipovska, A., Chia, T., Rajappa, L., Lingford, J.P., Razif, M.F., . . . Rackham, O. (2014) An artificial PPR scaffold for programmable RNA recognition. *Nature communications*, 5, 5729.

Crooks, G.E., Hon, G., Chandonia, J.M. and Brenner, S.E. (2004) WebLogo: a sequence logo generator. *Genome research*, **14**, 1188-1190.

de Longevialle, A.F., Small, I.D. and Lurin, C. (2010) Nuclearly Encoded Splicing Factors Implicated in RNA Splicing in Higher Plant Organelles. *Molecular plant*, **3**, 691-705.

Decatur, W.A. and Fournier, M.J. (2002) rRNA modifications and ribosome function. *Trends in Biochemical Sciences*, **27**, 344-351.

Delannoy, E., Le Ret, M., Faivre-Nitschke, E., Estavillo, G.M., Bergdoll, M., Taylor, N.L., . . . Gualberto, J.M. (2009) Arabidopsis tRNA adenosine deaminase arginine edits the wobble nucleotide of chloroplast tRNAArg(ACG) and is essential for efficient chloroplast translation. *The Plant cell*, **21**, 2058-2071.

Deng, X.-W. and Gruissem, W. (1987) Control of plastid gene expression during development: The limited role of transcriptional regulation. *Cell*, **49**, 379-387.

Deng, X.W., Tonkyn, J.C., Peter, G.F., Thornber, J.P. and Gruissem, W. (1989) Post-transcriptional control of plastid mRNA accumulation during adaptation of chloroplasts to different light quality environments. *The Plant cell*, **1**, 645-654.

Desai, U.J. and Pfaffle, P.K. (1995) Single-step purification of a thermostable DNA polymerase expressed in Escherichia coli. *BioTechniques*, **19**, 780-782, 784.

Driscoll, D.M., Wynne, J.K., Wallis, S.C. and Scott, J. (1989) An in vitro system for the editing of apolipoprotein B mRNA. *Cell*, **58**, 519-525.

Eberhard, S., Drapier, D. and Wollman, F.A. (2002) Searching limiting steps in the expression of chloroplast-encoded proteins: relations between gene copy number, transcription, transcript abundance and translation rate in the chloroplast of Chlamydomonas reinhardtii. *The Plant journal : for cell and molecular biology*, **31**, 149-160.

Edwards, K., Johnstone, C. and Thompson, C. (1991) A simple and rapid method for the preparation of plant genomic DNA for PCR analysis. *Nucleic acids research*, **19**, 1349.

El Baidouri, M., Kim, K.D., Abernathy, B., Arikit, S., Maumus, F., Panaud, O., . . . Jackson, S.A. (2015) A new approach for annotation of transposable elements using small RNA mapping. *Nucleic acids research*.

Falvey, A.K., Weiss, G.B., Krueger, L.J., Kantor, J.A. and Anderson, W.F. (1976) Transcription of single base oligonucleotides by ribonucleic acid-directed deoxyribonucleic acid polymerase. *Nucleic acids research*, **3**, 79-88.

Favory, J.J., Kobayshi, M., Tanaka, K., Peltier, G., Kreis, M., Valay, J.G. and Lerbs-Mache, S. (2005) Specific function of a plastid sigma factor for ndhF gene transcription. *Nucleic acids research*, **33**, 5991-5999.

Felder, S., Meierhoff, K., Sane, A.P., Meurer, J., Driemel, C., Plucken, H., . . . Westhoff, P. (2001) The nucleus-encoded HCF107 gene of Arabidopsis provides a link between intercistronic RNA processing and the accumulation of translation-competent psbH transcripts in chloroplasts. *The Plant cell*, **13**, 2127-2141.

Forner, J., Holzle, A., Jonietz, C., Thuss, S., Schwarzlander, M., Weber, B., . . . Binder, S. (2008) Mitochondrial mRNA polymorphisms in different Arabidopsis accessions. *Plant physiology*, **148**, 1106-1116.

Forner, J., Weber, B., Thuss, S., Wildum, S. and Binder, S. (2007) Mapping of mitochondrial mRNA termini in Arabidopsis thaliana: t-elements contribute to 5' and 3' end formation. *Nucleic acids research*, **35**, 3676-3692.

Fujii, S., Bond, C.S. and Small, I.D. (2011) Selection patterns on restorer-like genes reveal a conflict between nuclear and mitochondrial genomes throughout angiosperm evolution. *Proceedings of the National Academy of Sciences of the United States of America*, **108**, 1723-1728.

Georg, J., Dienst, D., Schurgers, N., Wallner, T., Kopp, D., Stazic, D., . . . Wilde, A. (2014) The small regulatory RNA SyR1/PsrR1 controls photosynthetic functions in cyanobacteria. *The Plant cell*, **26**, 3661-3679.

Germain, A., Herlich, S., Larom, S., Kim, S.H., Schuster, G. and Stern, D.B. (2011) Mutational analysis of Arabidopsis chloroplast polynucleotide phosphorylase reveals roles for both RNase PH core domains in polyadenylation, RNA 3'-end maturation and intron degradation. *The Plant journal : for cell and molecular biology*, **67**, 381-394.

Germain, A., Hotto, A.M., Barkan, A. and Stern, D.B. (2013) RNA processing and decay in plastids. *Wiley interdisciplinary reviews. RNA*, **4**, 295-316.

Germain, A., Kim, S.H., Gutierrez, R. and Stern, D.B. (2012) Ribonuclease II preserves chloroplast RNA homeostasis by increasing mRNA decay rates, and cooperates with polynucleotide phosphorylase in 3' end maturation. *The Plant journal : for cell and molecular biology*, **72**, 960-971.

Giege, P., Hoffmann, M., Binder, S. and Brennicke, A. (2000) RNA degradation buffers asymmetries of transcription in Arabidopsis mitochondria. *EMBO reports*, **1**, 164-170.

Gobert, A., Gutmann, B., Taschner, A., Gossringer, M., Holzmann, J., Hartmann, R.K., . . . Giege, P. (2010) A single Arabidopsis organellar protein has RNase P activity. *Nature structural & molecular biology*, **17**, 740-744.

Gualberto, J.M., Lamattina, L., Bonnard, G., Weil, J.H. and Grienenberger, J.M. (1989) RNA editing in wheat mitochondria results in the conservation of protein sequences. *Nature*, **341**, 660-662.

Gully, B.S., Cowieson, N., Stanley, W.A., Shearston, K., Small, I.D., Barkan, A. and Bond, C.S. (2015) The solution structure of the pentatricopeptide repeat protein PPR10 upon binding atpH RNA. *Nucleic acids research*.

Hafner, M., Landthaler, M., Burger, L., Khorshid, M., Hausser, J., Berninger, P., . . . Tuschl, T. (2010) Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP. *Cell*, **141**, 129-141.

Hafner, M., Renwick, N., Brown, M., Mihailovic, A., Holoch, D., Lin, C., . . . Tuschl, T. (2011) RNA-ligase-dependent biases in miRNA representation in deep-sequenced small RNA cDNA libraries. *Rna*, **17**, 1697-1712.

Haili, N., Arnal, N., Quadrado, M., Amiar, S., Tcherkez, G., Dahan, J., . . . Mireau, H. (2013) The pentatricopeptide repeat MTSF1 protein stabilizes the nad4 mRNA in Arabidopsis mitochondria. *Nucleic acids research*, **41**, 6650-6663.

Hajdukiewicz, P.T., Allison, L.A. and Maliga, P. (1997) The two RNA polymerases encoded by the nuclear and the plastid compartments transcribe distinct groups of genes in tobacco plastids. *The EMBO journal*, **16**, 4041-4048.

Hammani, K. and Barkan, A. (2014) An mTERF domain protein functions in group II intron splicing in maize chloroplasts. *Nucleic acids research*, **42**, 5033-5042.

Hammani, K., Bonnard, G., Bouchoucha, A., Gobert, A., Pinker, F., Salinas, T. and Giege, P. (2014) Helical repeats modular proteins are major players for organelle gene expression. *Biochimie*, **100**, 141-150.

Hammani, K., Cook, W.B. and Barkan, A. (2012) RNA binding and RNA remodeling activities of the half-a-tetratricopeptide (HAT) protein HCF107 underlie its effects on gene expression. *Proceedings of the National Academy of Sciences of the United States of America*, **109**, 5651-5656.

Hammani, K. and Giege, P. (2014) RNA metabolism in plant mitochondria. *Trends in plant science*, **19**, 380-389.

Hammani, K., Okuda, K., Tanz, S.K., Chateigner-Boutin, A.L., Shikanai, T. and Small, I. (2009) A study of new Arabidopsis chloroplast RNA editing mutants reveals general features of editing factors and their target sites. *The Plant cell*, **21**, 3686-3699.

Hashimoto, M., Endo, T., Peltier, G., Tasaka, M. and Shikanai, T. (2003) A nucleus-encoded factor, CRR2, is essential for the expression of chloroplast ndhB in Arabidopsis. *The Plant journal : for cell and molecular biology*, **36**, 541-549.

Hauler, A., Jonietz, C., Stoll, B., Stoll, K., Braun, H.P. and Binder, S. (2013) RNA Processing Factor 5 is required for efficient 5' cleavage at a processing site conserved in RNAs of three different mitochondrial genes in Arabidopsis thaliana. *The Plant journal : for cell and molecular biology*, **74**, 593-604.

Hayes, M.L., Giang, K., Berhane, B. and Mulligan, R.M. (2013) Identification of two pentatricopeptide repeat genes required for RNA editing and zinc binding by C-terminal cytidine deaminase-like domains. *The Journal of biological chemistry*, **288**, 36519-36529.

Hertel, S., Zoschke, R., Neumann, L., Qu, Y., Axmann, I.M. and Schmitz-Linneweber, C. (2013) Multiple checkpoints for the expression of the chloroplast-encoded splicing factor MatK. *Plant physiology*, **163**, 1686-1698.

Holec, S., Lange, H., Kuhn, K., Alioua, M., Borner, T. and Gagliardi, D. (2006) Relaxed transcription in Arabidopsis mitochondria is counterbalanced by RNA stability control mediated by polyadenylation and polynucleotide phosphorylase. *Molecular and cellular biology*, **26**, 2869-2876.

Holzle, A., Jonietz, C., Torjek, O., Altmann, T., Binder, S. and Forner, J. (2011) A RESTORER OF FERTILITY-like PPR gene is required for 5'-end processing of the nad4 mRNA in mitochondria of Arabidopsis thaliana. *The Plant journal : for cell and molecular biology*, **65**, 737-744.

Hotto, A.M., Castandet, B., Gilet, L., Higdon, A., Condon, C. and Stern, D.B. (2015) Arabidopsis Chloroplast Mini-Ribonuclease III Participates in rRNA Maturation and Intron Recycling. *The Plant Cell Online*.

Hotto, A.M., Schmitz, R.J., Fei, Z., Ecker, J.R. and Stern, D.B. (2011) Unexpected Diversity of Chloroplast Noncoding RNAs as Revealed by Deep Sequencing of the Arabidopsis Transcriptome. *G3*, **1**, 559-570.

Hricova, A., Quesada, V. and Micol, J.L. (2006) The SCABRA3 nuclear gene encodes the plastid RpoTp RNA polymerase, which is required for chloroplast biogenesis and mesophyll cell proliferation in Arabidopsis. *Plant physiology*, **141**, 942-956.

Iyer, L.M., Zhang, D., Rogozin, I.B. and Aravind, L. (2011) Evolution of the deaminase fold and multiple origins of eukaryotic editing and mutagenic nucleic acid deaminases from bacterial toxin systems. *Nucleic acids research*, **39**, 9473-9497.

Jacobs, J. and Kuck, U. (2011) Function of chloroplast RNA-binding proteins. *Cellular and molecular life sciences : CMLS*, **68**, 735-748.

Johnson, X., Wostrikoff, K., Finazzi, G., Kuras, R., Schwarz, C., Bujaldon, S., . . . Vallon, O. (2010) MRL1, a conserved Pentatricopeptide repeat protein, is required for stabilization of rbcL mRNA in Chlamydomonas and Arabidopsis. *The Plant cell*, **22**, 234-248.

Jonietz, C., Forner, J., Hildebrandt, T. and Binder, S. (2011) RNA PROCESSING FACTOR3 is crucial for the accumulation of mature ccmC transcripts in mitochondria of Arabidopsis accession Columbia. *Plant physiology*, **157**, 1430-1439.

Jonietz, C., Forner, J., Holzle, A., Thuss, S. and Binder, S. (2010) RNA PROCESSING FACTOR2 is required for 5' end processing of nad9 and cox3 mRNAs in mitochondria of Arabidopsis thaliana. *The Plant cell*, **22**, 443-453.

Karcher, D. and Bock, R. (1998) Site-selective inhibition of plastid RNA editing by heat shock and antibiotics: a role for plastid translation in RNA editing. *Nucleic acids research*, **26**, 1185-1190.

Karcher, D. and Bock, R. (2009) Identification of the chloroplast adenosine-to-inosine tRNA editing enzyme. *Rna*, **15**, 1251-1257.

Kasschau, K.D., Fahlgren, N., Chapman, E.J., Sullivan, C.M., Cumbie, J.S., Givan, S.A. and Carrington, J.C. (2007) Genome-wide profiling and analysis of Arabidopsis siRNAs. *PLoS biology*, 5, e57.

Kikuchi, S., Bedard, J., Hirano, M., Hirabayashi, Y., Oishi, M., Imai, M., . . . Nakai, M. (2013) Uncovering the protein translocon at the chloroplast inner envelope membrane. *Science*, **339**, 571-574.

Kindgren, P., Yap, A., Bond, C.S. and Small, I. (2015) Predictable Alteration of Sequence Recognition by RNA Editing Factors from Arabidopsis. *The Plant cell*.

Klaff, P. and Gruissem, W. (1991) Changes in Chloroplast mRNA Stability during Leaf Development. *The Plant cell*, **3**, 517-529.

Klein, R.R., Mason, H.S. and Mullet, J.E. (1988) Light-regulated translation of chloroplast proteins. I. Transcripts of psaA-psaB, psbA, and rbcL are associated with polysomes in dark-grown and illuminated barley seedlings. *The Journal of cell biology*, **106**, 289-301.

Konig, J., Zarnack, K., Rot, G., Curk, T., Kayikci, M., Zupan, B., . . . Ule, J. (2010) iCLIP reveals the function of hnRNP particles in splicing at individual nucleotide resolution. *Nature structural & molecular biology*, **17**, 909-915.

Kotera, E., Tasaka, M. and Shikanai, T. (2005) A pentatricopeptide repeat protein is essential for RNA editing in chloroplasts. *Nature*, **433**, 326-330.

Koussevitzky, S., Nott, A., Mockler, T.C., Hong, F., Sachetto-Martins, G., Surpin, M., . . . Chory, J. (2007) Signals from chloroplasts converge to regulate nuclear gene expression. *Science*, 316, 715-719.

Kuhn, K., Weihe, A. and Borner, T. (2005) Multiple promoters are a common feature of mitochondrial genes in Arabidopsis. *Nucleic acids research*, **33**, 337-346.

Kupsch, C., Ruwe, H., Gusewski, S., Tillich, M., Small, I. and Schmitz-Linneweber, C. (2012) Arabidopsis chloroplast RNA binding proteins CP31A and CP29A associate with large transcript pools and confer cold stress tolerance by influencing multiple chloroplast RNA processing steps. *The Plant cell*, **24**, 4266-4280.

Kuroda, H., Suzuki, H., Kusumegi, T., Hirose, T., Yukawa, Y. and Sugiura, M. (2007) Translation of psbC mRNAs starts from the downstream GUG, not the upstream AUG, and requires the extended Shine-Dalgarno sequence in tobacco chloroplasts. *Plant & cell physiology*, **48**, 1374-1378.

Lambowitz, A.M. and Zimmerly, S. (2004) MOBILE GROUP II INTRONS. *Annual review of genetics*, **38**, 1-35.

Lamesch, P., Berardini, T.Z., Li, D., Swarbreck, D., Wilks, C., Sasidharan, R., . . . Huala, E. (2012) The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools. *Nucleic acids research*, **40**, D1202-1210.

Langmead, B., Trapnell, C., Pop, M. and Salzberg, S.L. (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome biology*, **10**, R25.

Larkin, M.A., Blackshields, G., Brown, N.P., Chenna, R., McGettigan, P.A., McWilliam, H., . . . Higgins, D.G. (2007) Clustal W and Clustal X version 2.0. *Bioinformatics*, 23, 2947-2948.

Lawrence, M., Huber, W., Pages, H., Aboyoun, P., Carlson, M., Gentleman, R., . . . Carey, V.J. (2013) Software for computing and annotating genomic ranges. *PLoS computational biology*, **9**, e1003118.

Legen, J., Kemp, S., Krause, K., Profanter, B., Herrmann, R.G. and Maier, R.M. (2002) Comparative analysis of plastid transcription profiles of entire plastid chromosomes from tobacco attributed to wild-type and PEP-deficient transcription machineries. *The Plant journal : for cell and molecular biology*, **31**, 171-188.

Leister, D. (2003) Chloroplast research in the genomic age. *Trends in genetics : TIG*, **19**, 47-56.

Lerbs-Mache, S. (2011) Function of plastid sigma factors in higher plants: regulation of gene expression or just preservation of constitutive transcription? *Plant molecular biology*, **76**, 235-249.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., . . . Durbin, R. (2009a) The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25, 2078-2079.

Li, J.B., Levanon, E.Y., Yoon, J.K., Aach, J., Xie, B., Leproust, E., . . . Church, G.M. (2009b) Genome-wide identification of human RNA editing sites by parallel DNA capturing and sequencing. *Science*, **324**, 1210-1213.

Li, Q., Yan, C., Xu, H., Wang, Z., Long, J., Li, W., . . . Yan, N. (2014) Examination of the dimerization states of the single-stranded RNA recognition protein pentatricopeptide repeat 10 (PPR10). *The Journal of biological chemistry*, **289**, 31503-31512.

Liere, K., Weihe, A. and Borner, T. (2011) The transcription machineries of plant mitochondria and chloroplasts: Composition, function, and regulation. *Journal of plant physiology*, **168**, 1345-1360.

Lin, X., Kaul, S., Rounsley, S., Shea, T.P., Benito, M.I., Town, C.D., . . . Venter, J.C. (1999) Sequence and analysis of chromosome 2 of the plant Arabidopsis thaliana. *Nature*, **402**, 761-768.

Lisitsky, I. and Schuster, G. (1995) Phosphorylation of a chloroplast RNA-binding protein changes its affinity to RNA. *Nucleic acids research*, **23**, 2506-2511.

Liu, G., Mercer, T.R., Shearwood, A.M., Siira, S.J., Hibbs, M.E., Mattick, J.S., . . . Filipovska, A. (2013a) Mapping of mitochondrial RNA-protein interactions by digital RNase footprinting. *Cell reports*, **5**, 839-848.

Liu, S., Melonek, J., Boykin, L.M., Small, I. and Howell, K.A. (2013b) PPR-SMRs: ancient proteins with enigmatic functions. *RNA biology*, **10**, 1501-1510.

Liu, X., Yu, F. and Rodermel, S. (2010) An Arabidopsis pentatricopeptide repeat protein, SUPPRESSOR OF VARIEGATION7, is required for FtsH-mediated chloroplast biogenesis. *Plant physiology*, **154**, 1588-1601.

Loiselay, C., Gumpel, N.J., Girard-Bascou, J., Watson, A.T., Purton, S., Wollman, F.A. and Choquet, Y. (2008) Molecular identification and function of cis- and trans-acting determinants for petA transcript stability in Chlamydomonas reinhardtii chloroplasts. *Molecular and cellular biology*, **28**, 5529-5542.

Loizeau, K., Qu, Y., Depp, S., Fiechter, V., Ruwe, H., Lefebvre-Legendre, L., . . . Goldschmidt-Clermont, M. (2014) Small RNAs reveal two target sites of the RNA-maturation factor Mbb1 in the chloroplast of Chlamydomonas. *Nucleic acids research*, **42**, 3286-3297.

Lurin, C., Andres, C., Aubourg, S., Bellaoui, M., Bitton, F., Bruyere, C., . . . Small, I. (2004) Genome-wide analysis of Arabidopsis pentatricopeptide repeat proteins reveals their essential role in organelle biogenesis. *The Plant cell*, **16**, 2089-2103.

Maier, R.M., Hoch, B., Zeltz, P. and Kossel, H. (1992) Internal editing of the maize chloroplast ndhA transcript restores codons for conserved amino acids. *The Plant cell*, **4**, 609-616.

Majeran, W., Friso, G., Asakura, Y., Qu, X., Huang, M., Ponnala, L., . . . van Wijk, K.J. (2012) Nucleoid-enriched proteomes in developing plastids and chloroplasts from maize leaves: a new conceptual framework for nucleoid functions. *Plant physiology*, **158**, 156-189.

Malik Ghulam, M., Courtois, F., Lerbs-Mache, S. and Merendino, L. (2013) Complex processing patterns of mRNAs of the large ATP synthase operon in Arabidopsis chloroplasts. *PloS one*, **8**, e78265.

Manavski, N., Guyon, V., Meurer, J., Wienand, U. and Brettschneider, R. (2012) An essential pentatricopeptide repeat protein facilitates 5' maturation and translation initiation of rps3 mRNA in maize mitochondria. *The Plant cell*, **24**, 3087-3105.

Martin, M. (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. *2011*, **17**.

Maruyama, K., Sato, N. and Ohta, N. (1999) Conservation of structure and cold-regulation of RNA-binding proteins in cyanobacteria: probable convergent evolution with eukaryotic glycine-rich RNA-binding proteins. *Nucleic acids research*, **27**, 2029-2036.

Massenet, O., Martinez, P., Seyer, P. and Briat, J.-F. (1987) Sequence organization of the chloroplast ribosomal spacer ofSpinacia oleracea including the 3′ end of the 16S rRNA and the 5′ end of the 23S rRNA. *Plant molecular biology*, **10**, 53-63.

Meierhoff, K., Felder, S., Nakamura, T., Bechtold, N. and Schuster, G. (2003) HCF152, an Arabidopsis RNA binding pentatricopeptide repeat protein involved in the processing of chloroplast psbB-psbT-psbH-petB-petD RNAs. *The Plant cell*, **15**, 1480-1495.

Meister, G. (2013) Argonaute proteins: functional insights and emerging roles. *Nat Rev Genet*, **14**, 447-459.

Michalovova, M., Vyskot, B. and Kejnovsky, E. (2013) Analysis of plastid and mitochondrial DNA insertions in the nucleus (NUPTs and NUMTs) of six plant species: size, relative age and chromosomal localization. *Heredity*, **111**, 314-320.

Murashige, T. and Skoog, F. (1962) A Revised Medium for Rapid Growth and Bio Assays with Tobacco Tissue Cultures. *Physiologia Plantarum*, **15**, 473-497.

Nakamura, T., Ohta, M., Sugiura, M. and Sugita, M. (1999) Chloroplast ribonucleoproteins are associated with both mRNAs and intron-containing precursor tRNAs. *FEBS letters*, **460**, 437-441.

Nakamura, T., Ohta, M., Sugiura, M. and Sugita, M. (2001) Chloroplast ribonucleoproteins function as a stabilizing factor of ribosome-free mRNAs in the stroma. *The Journal of biological chemistry*, **276**, 147-152.

Nakamura, T., Schuster, G., Sugiura, M. and Sugita, M. (2004) Chloroplast RNA-binding and pentatricopeptide repeat proteins. *Biochemical Society transactions*, **32**, 571-574.

Nicol, J.W., Helt, G.A., Blanchard, S.G., Jr., Raja, A. and Loraine, A.E. (2009) The Integrated Genome Browser: free software for distribution and exploration of genome-scale datasets. *Bioinformatics*, **25**, 2730-2731.

Noordally, Z.B., Ishii, K., Atkins, K.A., Wetherill, S.J., Kusakina, J., Walton, E.J., . . . Dodd, A.N. (2013) Circadian control of chloroplast transcription by a nuclear-encoded timing signal. *Science*, **339**, 1316-1319.

Novoa, Eva M., Pavon-Eternod, M., Pan, T. and Ribas de Pouplana, L. (2012) A Role for tRNA Modifications in Genome Structure and Codon Usage. *Cell*, **149**, 202-213.

Okuda, K., Chateigner-Boutin, A.L., Nakamura, T., Delannoy, E., Sugita, M., Myouga, F., . . . Shikanai, T. (2009) Pentatricopeptide repeat proteins with the DYW motif have distinct molecular functions in RNA editing and RNA cleavage in Arabidopsis chloroplasts. *The Plant cell*, **21**, 146-156.

Okuda, K., Myouga, F., Motohashi, R., Shinozaki, K. and Shikanai, T. (2007) Conserved domain structure of pentatricopeptide repeat proteins involved in chloroplast RNA editing. *Proceedings of the National Academy of Sciences of the United States of America*, **104**, 8178-8183.

Okuda, K., Shoki, H., Arai, M., Shikanai, T., Small, I. and Nakamura, T. (2014) Quantitative analysis of motifs contributing to the interaction between PLS-subfamily members and their target RNA sequences in plastid RNA editing. *The Plant journal : for cell and molecular biology*, **80**, 870-882.

Ostersetzer, O., Cooke, A.M., Watkins, K.P. and Barkan, A. (2005) CRS1, a chloroplast group II intron splicing factor, promotes intron folding through specific interactions with two intron domains. *The Plant cell*, **17**, 241-255.

Pall, G.S., Codony-Servat, C., Byrne, J., Ritchie, L. and Hamilton, A. (2007) Carbodiimide-mediated cross-linking of RNA to nylon membranes improves the detection of siRNA, miRNA and piRNA by northern blot. *Nucleic acids research*, **35**, e60.

Pall, G.S. and Hamilton, A.J. (2008) Improved northern blot method for enhanced detection of small RNA. *Nature protocols*, **3**, 1077-1084.

Perrin, R., Meyer, E.H., Zaepfel, M., Kim, Y.J., Mache, R., Grienenberger, J.M., . . . Gagliardi, D. (2004) Two exoribonucleases act sequentially to process mature 3'-ends of atp9 mRNAs in Arabidopsis mitochondria. *The Journal of biological chemistry*, **279**, 25440-25446.

Pfalz, J., Bayraktar, O.A., Prikryl, J. and Barkan, A. (2009) Site-specific binding of a PPR protein defines and stabilizes 5' and 3' mRNA termini in chloroplasts. *The EMBO journal*, **28**, 2042-2052.

Pfitzinger, H., Weil, J.H., Pillay, D.T. and Guillemaut, P. (1990) Codon recognition mechanisms in plant chloroplasts. *Plant molecular biology*, **14**, 805-814.

Powikrowska, M., Oetke, S., Jensen, P.E. and Krupinska, K. (2014) Dynamic composition, shaping and organization of plastid nucleoids. *Frontiers in plant science*, **5**, 424.

Prikryl, J., Rojas, M., Schuster, G. and Barkan, A. (2011) Mechanism of RNA stabilization and translational activation by a pentatricopeptide repeat protein. *Proceedings of the National Academy of Sciences of the United States of America*, **108**, 415-420.

Pyke, K.A. (1999) Plastid division and development. *The Plant cell*, **11**, 549-556.

Quinlan, A.R. and Hall, I.M. (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, **26**, 841-842.

Rackham, O. and Filipovska, A. (2012) The role of mammalian PPR domain proteins in the regulation of mitochondrial gene expression. *Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms*, **1819**, 1008-1016.

Raczynska, K.D., Le Ret, M., Rurek, M., Bonnard, G., Augustyniak, H. and Gualberto, J.M. (2006) Plant mitochondrial genes can be expressed from mRNAs lacking stop codons. *FEBS letters*, **580**, 5641-5646.

Raina, M. and Ibba, M. (2014) tRNAs as regulators of biological processes. *Frontiers in genetics*, **5**, 171.

Rajagopalan, R., Vaucheret, H., Trejo, J. and Bartel, D.P. (2006) A diverse and evolutionarily fluid set of microRNAs in Arabidopsis thaliana. *Genes & development*, **20**, 3407-3425.

Raynaud, C., Loiselay, C., Wostrikoff, K., Kuras, R., Girard-Bascou, J., Wollman, F.A. and Choquet, Y. (2007) Evidence for regulatory function of nucleus-encoded factors on mRNA stabilization and translation in the chloroplast. *Proceedings of the National Academy of Sciences of the United States of America*, **104**, 9093-9098.

Reiland, S., Messerli, G., Baerenfaller, K., Gerrits, B., Endler, A., Grossmann, J., . . . Baginsky, S. (2009) Large-scale Arabidopsis phosphoproteome profiling reveals novel chloroplast kinase substrates and phosphorylation networks. *Plant physiology*, **150**, 889-903.

Rott, R., Liveanu, V., Drager, R.G., Stern, D.B. and Schuster, G. (1998) The sequence and structure of the 3'-untranslated regions of chloroplast transcripts are important determinants of mRNA accumulation and stability. *Plant molecular biology*, **36**, 307-314.

Ruwe, H. (2010) Die Rolle des chloroplastidären Ribonukleoproteins CP31A für die Prozessierung und Stabilisierung plastidärer Transkripte: Freie Universität Berlin.

Ruwe, H., Castandet, B., Schmitz-Linneweber, C. and Stern, D.B. (2013) Arabidopsis chloroplast quantitative editotype. *FEBS letters*, **587**, 1429-1433.

Ruwe, H., Kupsch, C., Teubner, M. and Schmitz-Linneweber, C. (2011) The RNA-recognition motif in chloroplasts. *Journal of plant physiology*, **168**, 1361-1371.

Ruwe, H. and Schmitz-Linneweber, C. (2012) Short non-coding RNA fragments accumulating in chloroplasts: footprints of RNA binding proteins? *Nucleic acids research*, **40**, 3106-3116.

Salone, V., Rudinger, M., Polsakiewicz, M., Hoffmann, B., Groth-Malonek, M., Szurek, B., . . . Lurin, C. (2007) A hypothesis on the identification of the editing enzyme in plant organelles. *FEBS letters*, **581**, 4132-4138.

Sambrook, J. and Russell, D.W. (2001) *Molecular Cloning: A Laboratory Manual*: Cold Spring Harbor Laboratory Press.

Sane, A.P., Stein, B. and Westhoff, P. (2005) The nuclear gene HCF107 encodes a membrane-associated R-TPR (RNA tetratricopeptide repeat)-containing protein involved in expression of the plastidial psbH gene in Arabidopsis. *The Plant journal : for cell and molecular biology*, **42**, 720-730.

Scharff, L.B., Childs, L., Walther, D. and Bock, R. (2011) Local absence of secondary structure permits translation of mRNAs that lack ribosome-binding sites. *PLoS genetics*, **7**, e1002155.

Schmitz-Linneweber, C. and Small, I. (2008) Pentatricopeptide repeat proteins: a socket set for organelle gene expression. *Trends in plant science*, **13**, 663-670.

Schmitz-Linneweber, C., Williams-Carrier, R.E., Williams-Voelker, P.M., Kroeger, T.S., Vichas, A. and Barkan, A. (2006) A pentatricopeptide repeat protein facilitates the trans-splicing of the maize chloroplast rps12 pre-mRNA. *The Plant cell*, **18**, 2650-2663.

Schmitz, R.J., Schultz, M.D., Lewsey, M.G., O'Malley, R.C., Urich, M.A., Libiger, O., . . . Ecker, J.R. (2011) Transgenerational epigenetic instability is a source of novel methylation variants. *Science*, **334**, 369-373.

Schwarz, C., Elles, I., Kortmann, J., Piotrowski, M. and Nickelsen, J. (2007) Synthesis of the D2 protein of photosystem II in Chlamydomonas is controlled by a high molecular mass complex containing the RNA stabilization factor Nac2 and the translational activator RBP40. *The Plant cell*, **19**, 3627-3639.

Sharwood, R.E., Halpert, M., Luro, S., Schuster, G. and Stern, D.B. (2011) Chloroplast RNase J compensates for inefficient transcription termination by removal of antisense RNA. *Rna*, **17**, 2165-2176.

Shikanai, T. (2015) RNA editing in plants: Machinery and flexibility of site recognition. *Biochimica et biophysica acta*.

Silverman, I.M., Li, F., Alexander, A., Goff, L., Trapnell, C., Rinn, J.L. and Gregory, B.D. (2014) RNase-mediated protein footprint sequencing reveals protein-binding sites throughout the human transcriptome. *Genome biology*, **15**, R3.

Small, I.D. and Peeters, N. (2000) The PPR motif - a TPR-related motif prevalent in plant organellar proteins. *Trends Biochem Sci*, **25**, 46-47.

Small, I.D., Rackham, O. and Filipovska, A. (2013) Organelle transcriptomes: products of a deconstructed genome. *Current opinion in microbiology*, **16**, 652-658.

Steglich, C., Futschik, M.E., Lindell, D., Voss, B., Chisholm, S.W. and Hess, W.R. (2008) The challenge of regulation in a minimal photoautotroph: non-coding RNAs in Prochlorococcus. *PLoS genetics*, **4**, e1000173.

Stern, D.B., Goldschmidt-Clermont, M. and Hanson, M.R. (2010) Chloroplast RNA metabolism. *Annual review of plant biology*, **61**, 125-155.

Stern, D.B. and Gruissem, W. (1987) Control of plastid gene expression: 3' inverted repeats act as mRNA processing and stabilizing elements, but do not terminate transcription. *Cell*, **51**, 1145-1157.

Stoll, B., Zendler, D. and Binder, S. (2014) RNA processing factor 7 and polynucleotide phosphorylase are necessary for processing and stability of nad2 mRNA in Arabidopsis mitochondria. *RNA biology*, **11**, 968-976.

Stoll, K., Jonietz, C. and Binder, S. (2015) In Arabidopsis thaliana two co-adapted cyto-nuclear systems correlate with distinct ccmC transcript sizes. *The Plant journal : for cell and molecular biology*, **81**, 247-257.

Stoppel, R., Lezhneva, L., Schwenkert, S., Torabi, S., Felder, S., Meierhoff, K., . . . Meurer, J. (2011) Recruitment of a ribosomal release factor for light- and stress-dependent regulation of petB transcript stability in Arabidopsis chloroplasts. *The Plant cell*, **23**, 2680-2695.

Sun, T., Germain, A., Giloteaux, L., Hammani, K., Barkan, A., Hanson, M.R. and Bentolila, S. (2013) An RNA recognition motif-containing protein is required for plastid RNA editing in Arabidopsis and maize. *Proceedings of the National Academy of Sciences of the United States of America*, **110**, E1169-1178.

Takenaka, M. (2014) How complex are the editosomes in plant organelles? *Molecular plant*, **7**, 582-585.

Takenaka, M., Zehrmann, A., Brennicke, A. and Graichen, K. (2013a) Improved computational target site prediction for pentatricopeptide repeat RNA editing factors. *PloS one*, **8**, e65343.

Takenaka, M., Zehrmann, A., Verbitskiy, D., Hartel, B. and Brennicke, A. (2013b) RNA editing in plants and its evolution. *Annual review of genetics*, **47**, 335-352.

Takenaka, M., Zehrmann, A., Verbitskiy, D., Kugelmann, M., Hartel, B. and Brennicke, A. (2012) Multiple organellar RNA editing factor (MORF) family proteins are required for RNA editing in mitochondria and plastids of plants. *Proceedings of the National Academy of Sciences of the United States of America*, **109**, 5104-5109.

Tillich, M., Hardel, S.L., Kupsch, C., Armbruster, U., Delannoy, E., Gualberto, J.M., . . . Schmitz-Linneweber, C. (2009) Chloroplast ribonucleoprotein CP31A is required for editing and stability of specific chloroplast mRNAs. *Proceedings of the National Academy of Sciences of the United States of America*, **106**, 6002-6007.

Timmis, J.N., Ayliffe, M.A., Huang, C.Y. and Martin, W. (2004) Endosymbiotic gene transfer: organelle genomes forge eukaryotic chromosomes. *Nat Rev Genet*, **5**, 123-135.

Tokuhisa, J.G., Vijayan, P., Feldmann, K.A. and Browse, J.A. (1998) Chloroplast development at low temperatures requires a homolog of DIM1, a yeast gene encoding the 18S rRNA dimethylase. *The Plant cell*, **10**, 699-711.

Vaistij, F.E., Boudreau, E., Lemaire, S.D., Goldschmidt-Clermont, M. and Rochaix, J.D. (2000) Characterization of Mbb1, a nucleus-encoded tetratricopeptide-like repeat protein required for expression of the chloroplast psbB/psbT/psbH gene cluster in Chlamydomonas reinhardtii. *Proceedings of the National Academy of Sciences of the United States of America*, **97**, 14813-14818.

Voelker, R. and Barkan, A. (1995) Two nuclear mutations disrupt distinct pathways for targeting proteins to the chloroplast thylakoid. *The EMBO journal*, **14**, 3905-3914.

Voinnet, O. (2009) Origin, biogenesis, and activity of plant microRNAs. *Cell*, **136**, 669-687.

Wagoner, J.A., Sun, T., Lin, L. and Hanson, M.R. (2015) Cytidine Deaminase Motifs within the DYW Domain of Two Pentatricopeptide Repeat-containing Proteins Are Required for Site-specific Chloroplast RNA Editing. *The Journal of biological chemistry*, **290**, 2957-2968.

Walter, M., Kilian, J. and Kudla, J. (2002) PNPase activity determines the efficiency of mRNA 3'-end processing, the degradation of tRNA and the extent of polyadenylation in chloroplasts. *The EMBO journal*, **21**, 6905-6914.

Wang, L., Yu, X., Wang, H., Lu, Y.Z., de Ruiter, M., Prins, M. and He, Y.K. (2011) A novel class of heat-responsive small RNAs derived from the chloroplast genome of Chinese cabbage (Brassica rapa). *BMC genomics*, **12**, 289.

Watkins, K.P., Kroeger, T.S., Cooke, A.M., Williams-Carrier, R.E., Friso, G., Belcher, S.E., . . . Barkan, A. (2007) A ribonuclease III domain protein functions in group II intron splicing in maize chloroplasts. *The Plant cell*, **19**, 2606-2623.

Yagi, Y., Hayashi, S., Kobayashi, K., Hirayama, T. and Nakamura, T. (2013) Elucidation of the RNA recognition code for pentatricopeptide repeat proteins involved in organelle RNA editing in plants. *PloS one*, **8**, e57286.

Yamaguchi, K. and Subramanian, A.R. (2003) Proteomic identification of all plastid-specific ribosomal proteins in higher plant chloroplast 30S ribosomal subunit. *European Journal of Biochemistry*, **270**, 190-205.

Yamaguchi, K., von Knoblauch, K. and Subramanian, A.R. (2000) The Plastid Ribosomal Proteins: IDENTIFICATION OF ALL THE PROTEINS IN THE 30 S SUBUNIT OF AN ORGANELLE RIBOSOME (CHLOROPLAST). *Journal of Biological Chemistry*, **275**, 28455-28465.

Yap, A., Kindgren, P., Colas des Francs-Small, C., Kazama, T., Tanz, S.K., Toriyama, K. and Small, I. (2015) AEF1/MPR25 is implicated in RNA editing of plastid atpF and mitochondrial nad5 and also promotes atpF splicing in Arabidopsis and rice. *The Plant journal : for cell and molecular biology*.

Yin, P., Li, Q., Yan, C., Liu, Y., Liu, J., Yu, F., . . . Yan, N. (2013) Structural basis for the modular recognition of single-stranded RNA by PPR proteins. *Nature*, **504**, 168-171.

Zhang, X., Yazaki, J., Sundaresan, A., Cokus, S., Chan, S.W., Chen, H., . . . Ecker, J.R. (2006) Genome-wide high-resolution mapping and functional analysis of DNA methylation in arabidopsis. *Cell*, **126**, 1189-1201.

Zhelyazkova, P., Hammani, K., Rojas, M., Voelker, R., Vargas-Suarez, M., Borner, T. and Barkan, A. (2012a) Protein-mediated protection as the predominant mechanism for defining processed mRNA termini in land plant chloroplasts. *Nucleic acids research*, **40**, 3092-3105.

Zhelyazkova, P., Sharma, C.M., Forstner, K.U., Liere, K., Vogel, J. and Borner, T. (2012b) The primary transcriptome of barley chloroplasts: numerous noncoding RNAs and the dominating role of the plastid-encoded RNA polymerase. *The Plant cell*, **24**, 123-136.

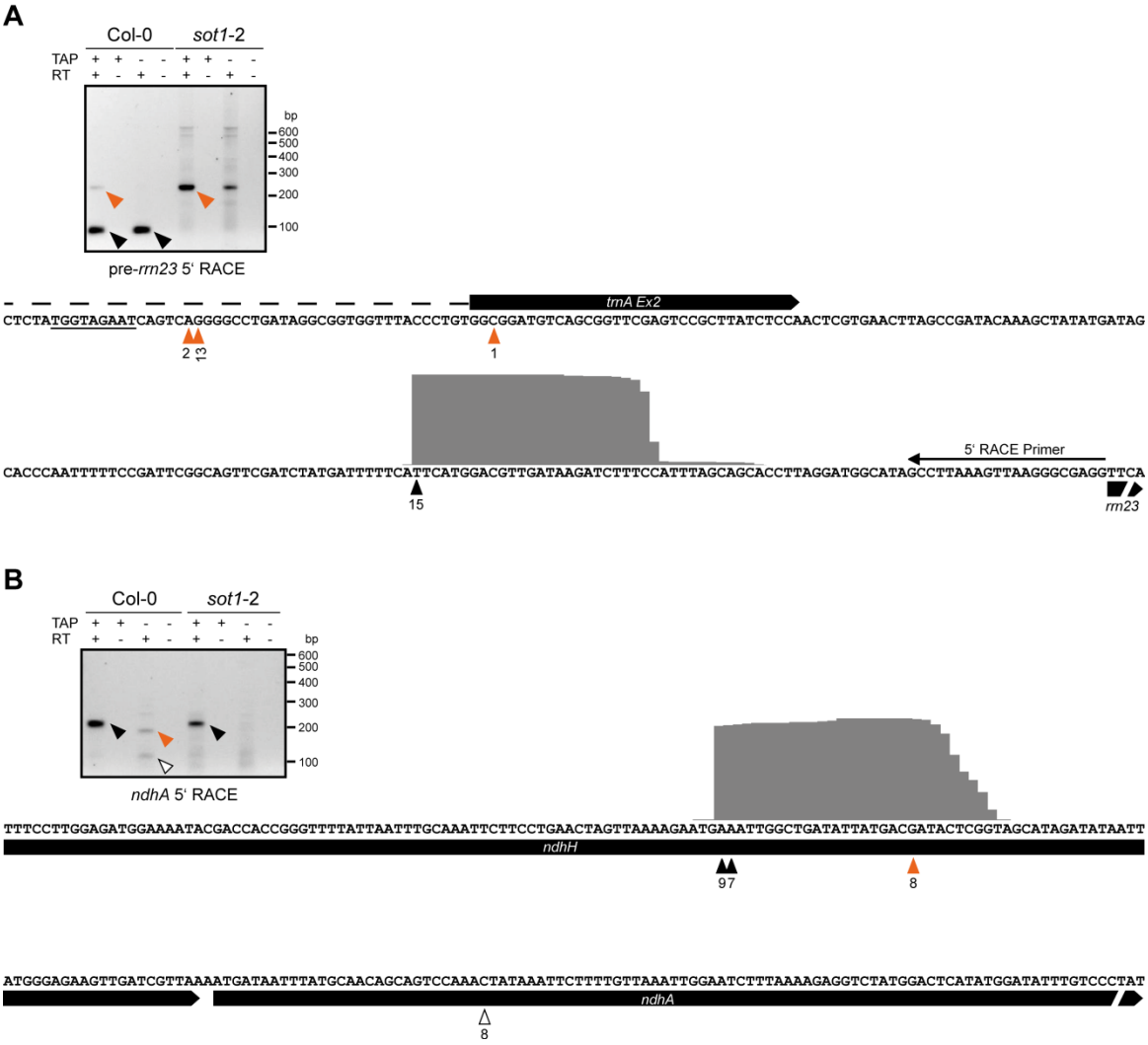
Zimorski, V., Ku, C., Martin, W.F. and Gould, S.B. (2014) Endosymbiotic theory for organelle origins. *Current opinion in microbiology*, **22C**, 38-48.

Zoschke, R., Nakamura, M., Liere, K., Sugiura, M., Borner, T. and Schmitz-Linneweber, C. (2010) An organellar maturase associates with multiple group II introns. *Proceedings of the National Academy of Sciences of the United States of America*, **107**, 3245-3250.

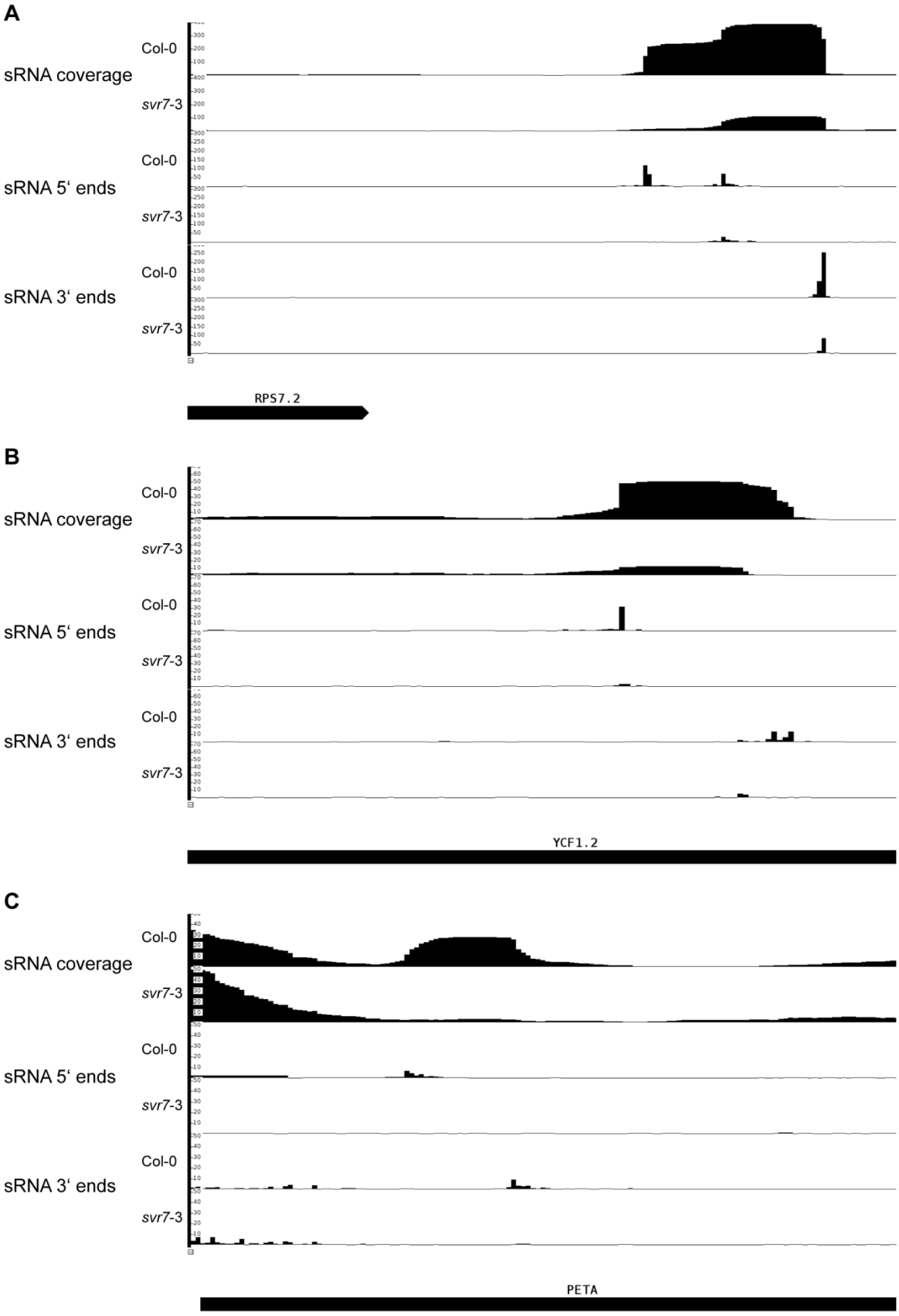
Zoschke, R., Qu, Y., Zubo, Y.O., Borner, T. and Schmitz-Linneweber, C. (2013) Mutation of the pentatricopeptide repeat-SMR protein SVR7 impairs accumulation and translation of chloroplast ATP synthase subunits in Arabidopsis thaliana. *Journal of plant research*, **126**, 403-414.

Zurawski, G., Perrot, B., Bottomley, W. and Whitfeld, P.R. (1981) The structure of the gene for the large subunit of ribulose 1,5-bisphosphate carboxylase from spinach chloroplast DNA. *Nucleic acids research*, **9**, 3251-3270.

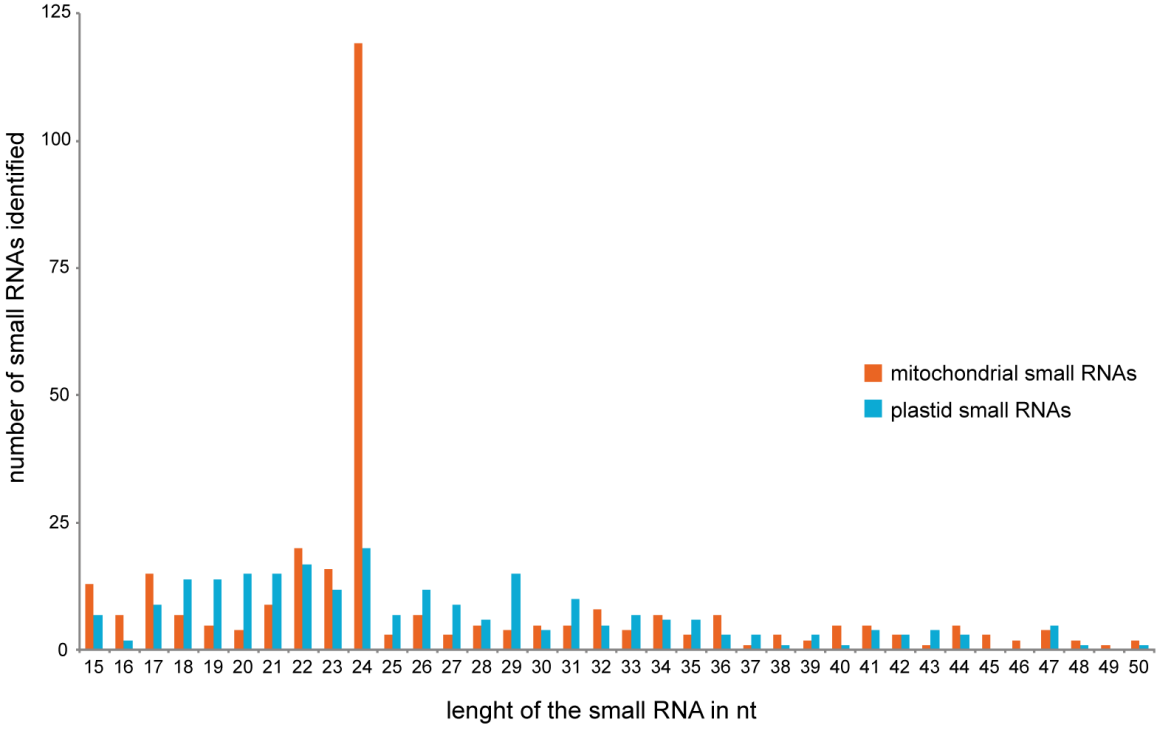
Appendix



Supplementary Figure 1: Transcript end mapping in *sot1*-2 mutants. 5' rapid amplification of cDNA ends analysis for precursors of *rrn23* (A) and *ndhA* transcripts (B). 1µg RNA from wild-type and *sot1*-2 mutant was treated with tobacco acid pyrophosphatase (TAP) to convert primary triphosphate ends into monophosphate. A control reaction was incubated without TAP. RNA was ligated to an adapter RNA oligonucleotide and reverse transcribed (+RT reactions). PCR products obtained with a gene-specific and adapter specific primer were separated on agarose gels (top of each panel). Bands marked with arrowheads were gel-purified and cloned. Positions of 5' ends determined by sequencing are shown in the lower part of the panel. Numbers below arrowheads indicate numbers of clones obtained at this position. The color indicates the origin of the clones from bands marked in the gel image. A -10 promoter element upstream of a primary 5' end for the *rrn23* gene is underlined.


A
Col-0
sRNA coverage
svr7-3
Col-0
sRNA 5‘ ends
svr7-3
Col-0
sRNA 3‘ ends
svr7-3
RPS7.2
B
Col-0
sRNA coverage
svr7-3
Col-0
sRNA 5‘ ends
svr7-3
Col-0
sRNA 3‘ ends
svr7-3
YCF1.2
C
Col-0
sRNA coverage
svr7-3
Col-0
sRNA 5‘ ends
svr7-3
Col-0
sRNA 3‘ ends
svr7-3
PETA

Supplementary Figure 2: Differential accumulation of small RNAs in *svr7*-3 mutants. Small RNA accumulation is shown in wild-type (Col-0) and the s*vr7*-3 mutant in three genomic regions that showed differential coverage (Figure 10) using the Integrated Genome Browser (Nicol et al. 2009). Small RNA coverage as well as abundance of small RNA 5' and 3' ends is shown with nucleotide resolution. The y-axis represents reads per million mapped chloroplast reads and is thus normalized. (A) A small RNA downstream of *rps7* shows strongly reduced coverage in the 5' region, whereas the 3' region of the small RNA accumulates almost normally. The small RNA is found at a 3' end of *rps7* (Figure 8) (B) The small RNAs found at the mature 3' end of *ndhF* and at the 3' end of *ycf1*as transcripts (Figure 17, Figure 20) are shortened at the 3' end and overall reduced. (C) A low abundance small RNA in the coding region of *petA* is lacking in *svr7*-3 mutants.



Supplementary Figure 3: Length distribution of small RNAs identified in plastids and mitochondria. Length of small RNAs identified using the algorithm described in 4.2.19. Counts of mitochondrial small RNAs are shown in orange, counts for plastid small RNAs in blue.

Supplementary Table 1: Prediction of PLS-PPRs for novel RNA editing sites. High scores (blue) indicate good overlap with the proposed PPR code (Yap et al. 2015).

	atpH 3'UTR	*ycf3* Intron 2	*rps4* 3'UTR	*ndhK*	*ndhK-ndhJ*	*rps18* 3'UTR	*ycf2* as	*ndhB* 3'UTR	*ndhB*	*ndhB*	genetically identified targets				
	13210	43350	45095	49849	49209	68453	91535	94622	96439	96457					
											atpF-12707				
AEF1	1.58	-2.57	-0.57	-0.21	-2.87	-2.01	0.68	1.22	0.05	-1.07	3.99				
											clpP-69942	*rpoA-78691*			
CLB19	-0.33	3.84	-1.69	-1.71	0.75	0.36	1.52	-2.69	-1.36	-1.35	1.96	1.98			
											psbE-64109				
CREF3	-0.30	0.53	-0.91	-1.42	-1.21	-0.30	0.10	-0.13	1.52	-0.19	1.68				
											ndhB-95225				
CREF7	0.07	-0.98	1.23	1.44	-1.70	-0.95	-1.88	0.97	-1.49	0.52	2.37				
											ndhD-116785				
CRR21	0.46	-0.66	-1.82	0.51	-1.70	-0.26	-2.51	-2.10	-2.20	-0.55	4.61				
											rpoB-25779	*ndhB-96419*	*ndhD-116281*		
CRR22	1.34	-0.80	-0.49	0.00	-0.48	3.18	0.62	-1.51	-1.91	-1.25	2.56	1.82	2.09		
											ndhB-96698	*ndhD-116290*			
CRR28	0.82	-0.16	0.01	-0.56	-1.45	2.05	-0.14	0.42	-0.65	-0.48	0.92	0.75			
											ndhD-117166				
CRR4	1.99	-0.67	1.88	0.70	-1.42	1.90	-0.49	0.87	-2.71	-0.56	4.81				
											ndhB-95650				
ELI1	0.04	-0.72	-0.68	-1.02	1.65	0.98	0.22	-1.18	-1.05	-0.02	2.45				
											rpoC1-21806				
FLV/DOT4	-0.53	-1.48	1.58	-2.82	-1.54	0.30	-0.64	1.14	0.42	1.76	3.14				
											psbF-63985				
LPA66	1.10	-1.43	0.18	2.04	-1.30	-1.00	0.97	0.61	-1.28	0.42	1.32				
											rpl23-86055				
OTP80	-1.40	-0.67	-1.19	-2.19	-1.87	-0.26	-1.48	-0.72	2.29	-0.78	2.69				
											matK-2930	*rpoB-23898*	*accD-58642*	*rps12-69553*	*ndhB-95608*
OTP81/QED1	0.86	-3.74	2.12	1.15	1.03	-0.16	-1.41	3.61	0.13	0.87	1.6	3	2.58	3.28	1.24
											ndhB-95644	*ndhG-118858*			
OTP82	1.07	1.62	-1.03	-0.52	-0.63	-0.35	2.29	0.19	-0.11	-0.19	0.4	3.52			
											psbZ-35800	*ndhB-94999*	*ndhF-112349*		
OTP84	1.77	-0.84	0.64	-0.24	-1.77	1.54	3.26	-0.13	-1.18	0.13	3.76	3.47	4.6		
											ndhD-116494				
OTP85	0.53	-0.52	-0.47	0.92	0.24	-0.26	1.08	0.84	-1.70	-1.04	2.16				
											rps14-37161				
OTP86	-0.27	-1.16	-1.16	-3.17	-1.61	-2.02	0.75	-0.64	1.03	1.30	0.78				
											accD-57868				
RARE1	-1.34	-2.04	-0.22	-0.41	0.64	-2.04	-1.22	0.51	-0.18	0.05	3.31				
											rpoB-25992				
YS1	-0.48	-3.98	-0.99	-2.37	-0.42	-2.37	-1.16	-0.82	-1.24	-1.78	3.57				

Supplementary Table 2: Small RNAs identified in chloroplasts

Name	Start	End	+/-	Start	End	+/-	Sequence
C1	1732	1758	+				CGAACCCGGAACTAGTCGGATGGAGT
C2	5782	5805	+				TCATACAAACGCTTGATTCACGC
C3	6054	6073	+				TGTCGAGCCAAGAGCACCT
C4	6637	6654	+				TCCGAATAGCGGGACCA
C5	7768	7790	+				CCATCAAAAGGAGAAGGGGAAA
C6	8402	8420	+				TTTTAATAGCCTGGCCTG
C7	8645	8663	+				GCGGGTATAGTTTAGTGG
C8	9589	9629	+				GCGTCCATTGTCTAATGGATAGGACATAGGTCTTCTAAAC
C9	9629	9664	+				CTTTGGTATAGGTTCAAATCCTATTGGACGCAATA
C10	14524	14548	+				TCGAATGAATTCAAGGACAAATTC
C11	17464	17488	+				TCTTATGAAATCTTGAATCAAACC
C12	24005	24023	+				GTTTCTTTTGAAGTCGAT
C13	24221	24245	+				CCTTGGGTTGTCACATGCGTCTGA
C14	24686	24712	+				GAACTCATTAAAGCTCGATTCGCATC
C15	26037	26060	+				TTTGTTCTTGCATATTCCTACTG
C16	27372	27406	+				GGCGGCATGGCCGAGTGGTAAGGCGGGGGACTGC
C17	30938	30957	+				ATCCAAGAAAGTCAGCCAG
C18	31368	31401	+				GCCCTTTTAACTCAGTGGTAGAGTAACGCCATG
C19	31418	31439	+				GGTTCAAATCCGATAAGGGGC
C20	32524	32548	+				TTTTGATCTTCGAAACCAATTAAA
C21	33710	33738	+				ATCAGCCTCATGAAAACCTTATATTCCC
C22	36489	36528	+				GCGGATATAGTCGAATGGTAAAATTTCTCCTTGCCAAGG
C23	36702	36730	+				GTTGCGGAGACAGGATTTGAACCCGTGA
C24	42061	42085	+				ATCCATAGGGTGCTCAACGGACCC
C25	44826	44853	+				GGAGAGATGGCCGAGTGGTTGAAGGCG
C26	46750	46783	+				GATTAGACTAAATCAATATGGATGGAGCTCAAA
C27	46872	46919	+				TCATAATGAGATCCTAAAAAAGGGGATATGGCGGAATTGGTAGACGC
C28	48174	48207	+				GCCGGGATAGCTCAGTTGGTAGAGCAGAGGACT
C29	52055	52096	+				ACCTACTTAACTCAGTGGTTAGAGTATTGCTTTCATACGGC
C30	54888	54920	+				TGTCGAGTAGACCTTGTTGTTTTGTTTTATTG
C31	56713	56756	+				GCATGTTGTCCTTTTCTTTTCATTCCGTATTGGAATAAAAAAA
C32	57014	57045	+				ATTGAATGACTATTCATCTATTGTTATTGTA
C33	57787	57806	+				GTATAAGAAAGTCAAAATG
C34	60672	60696	+				CGATAGAAATATTAGATCTAATAG
C35	61614	61643	+				GCTAACTTTATTGTAGAAATTTTCGGGAT
C36	65653	65684	+				TTAGGGAAGTACTTTAAGAAACATATGTATA
C37	67142	67163	+				GTAGAATAAATTAGAAAAGGT
C38	68106	68128	+				TTGCTATAAAACAAGCTCGTAT
C39	68252	68274	+				TCTGAAGGAATTAAAAAAGAGA
C40	68431	68457	+				TAATTTCTACTCTACCTTCCCCGAGC
C41	68512	68537	+				TATTTTTTTATGTCATTCGAAATTG
C42	74440	74470	+				GGTATACAAAGTCAACAGATCGTAATGAAT
C43	74793	74813	+				GGTAGTTCGACCGCGAAATT
C44	76626	76649	+				CTTTTCTATGATCGTACCCGACG
C45	78361	78383	+				TGAATACAGCATCGATAGGATA
C46	79542	79563	+				CCTCCTGCGGATTAGTCGACA
C47	82581	82603	+				ATAGGTAAGTTCTTTTTTCTTT
C48	84168	84211	+				TGGATGCCCGGGACCAAGTTATTATGATTTCTTTTTCCGCCTT
C49	84779	84808	+	153840	153869	-	ATTCTTCTTTTTGATCAATCAAAACCCCT
C50	85980	86000	+	152648	152668	-	GTCGATGACTATTCATAGCT
C51	86175	86199	+	152449	152473	-	CAATAAGAATGCTAGTTCTTACTG
C52	86848	86866	+	151782	151800	-	TCTTTTGGGTCTTGCAAT
C53	90772	90788	+	147860	147876	-	CGGGGTTCTGGCGGCA
C54	93567	93588	+	145060	145081	-	AACAAGAATTCTTGAACAGCG
C55	95346	95367	+	143281	143302	-	GTTCCGGTACGTAGACCAAAT
C56	96190	96205	+	142443	142458	-	GCAAAATGGATCCGT

Name	Start	End	+/-	Start	End	+/-	Sequence
C57	96823	96858	+	141790	141825	-	AGAGGAATACAGAGAGTTGAACATAGTAAAATAAG
C58	97642	97673	+	140975	141006	-	AATGGCAAGTGCTTTTCCTTGCGTGGATCCT
C59	98331	98354	+	140294	140317	-	CACGGACAAAGTCAGGGAAAACC
C60	99734	99754	+	138894	138914	-	GTAGCAACGGAACCGGGGAA
C61	100708	100739	+	137909	137940	-	AGGGATATAACTCAGCGGTAGAGTGTCACCT
C62	100821	100838	+	137810	137827	-	CGCTGTGATCGAATAAG
C63	100982	101002	+	137646	137666	-	AAGGAAGCTATAAGTAATGC
C64	101011	101036	+	137612	137637	-	TCTCATGGAGAGTTCGATCCTGGCT
C65	102837	102856	+	135792	135811	-	TTGCGTCGTTGTGCCTGGG
C66	103049	103064	+	135584	135599	-	TCGTGGGATCCGGGC
C67	103664	103702	+	134946	134984	-	GGGGATATAGCTCAGTTGGTAGAGCTCCGCTCTTGCAA
C68	103702	103744	+	134904	134946	-	TTGGGTCGTTGCGATTACGGGTTGGGTGTCTAATTGTCCAGG
C69	104617	104643	+	134005	134031	-	TTCATGGACGTTGATAAGATCTTTCC
C70	104691	104715	+	133933	133957	-	TCAAACGAGGAAAGGCTTACGGTG
C71	106446	106469	+	132179	132202	-	GGGGGTCGCAGTGACCAGGCCCG
C72	107948	107982	+	130666	130700	-	TATTCTGGTGTCCTAGGCGTAGAGGAACAACACC
C73	108301	108320	+	130328	130347	-	GGGCTTGTAGCTCAGAGGA
C74	114269	114296	+				GCCGCTATGGTGAAATTGGTAGACACG
C75	117018	117036	+				GCGTAGGTCGTTAGAAGA
C76	120425	120451	+				ACATGAGGTCTTGGCCTCATACGGCT
C77	124504	124541	+				TACCGCTATTTCGTTTGGATTGTTTAGTCTAACCAAG
C78	127796	127825	+				TTAGGTAAATATTCTTTTTTAGCTTCGTT
C79	128700	128725	+	109923	109948	-	GACCAATTAACCAACCAACAAAACT
C80	128779	128802	+	109846	109869	-	TCTGGCTAACATTGAACTTGGTA
C81	129275	129304	+	109344	109373	-	TCTGGATTATTATATGATGATTTTGCAAC
C82	129417	129432	+	109216	109231	-	AGAGCCGCTTTGAGG
C83	129564	129596	+	109052	109084	-	TCCTCAGTAGCTCAGTGGTAGAGCGGTCGGCT
C84	130495	130512	+	108136	108153	-	TTTGAATAAGACAACCT
C85	132125	132145	+	106503	106523	-	CCATACATGGTCTTACGACT
C86	135063	135086	+	103562	103585	-	TGAACCAGAGACCTCGCCCGTGA
C87	138210	138238	+	100410	100438	-	GACTCGGCATGTTCTATTCGATACGGGT
C88	138961	139008	+	99640	99687	-	CAACATAGGTCGTCGAAAGGATCTCGGAGACCCGCCAAAGCACGAAA
C89	141235	141276	+	97372	97413	-	TCAATAGAAAAAGAAAAAATCGGAATTGATCGATCTCTTTC
C90	141472	141494	+	97154	97176	-	AGTTACTAATTCATGATCTGGC
C91	142003	142021	+	96627	96645	-	ATACGATCTAATGAGGCT
C92	142232	142261	+	96387	96416	-	ATCAATGGACTCCTGACGTATACGAAGGA
C93	144292	144333	+	94315	94356	-	GCCTTGGTGGTGAAATGGTAGACACGCGAGACTCAAAATCT
C94	145050	145084	+	93564	93598	-	TCCGGTTGTTCGCTGTTCAAGAATTCTTGTTTAG
C95	147249	147268	+	91380	91399	-	CCTAGAGGGGGATAGGGCT
C96	148366	148410	+	90238	90282	-	TCTGAAAAAGTATCTAAAAATATCAAATTTAGATATTTGTACCC
C97	150261	150295	+	88353	88387	-	AAAGGCAAATCCCTTATGATACACCAGATCCGGC
C98	150827	150852	+	87796	87821	-	CTGATTCTATCTCTCTTCCTTCCGT
C99	152048	152072	+	86576	86600	-	AAATATGAATGAAAGATCCCACTG
C100	152263	152278	+	86370	86385	-	GCATCCATGGCTGAA
C101	152337	152380	+	86268	86311	-	CGCCAATCGGACCCTCCAATAAGTCTATTGGAATTGGCTCTGT
C102	152749	152771	+	85877	85899	-	GTTATTCTATTCCACCTCTTAG
C103	154205	154236	+	84412	84443	-	GTAGAAAAAAACCCGTAACCCCCTGGGGTTA
C104	770	799	+				CTGATCAAACTAGAAGTTACCAAGGAACC
C105	6146	6177	+				ACGTTGCTTTCTACCACATCGTTTTAAACGA
C106	7472	7493	+				ACAAATAACTTTCTGAAACCT
C107	8933	8959	+				GAAAAGTGTCTTTCTAATCGTAACTA
C108	9556	9574	+				TTTTAACAATAGGAAAGT
C109	14185	14206	+				ATTTCCGAAAAGTCGAAAACT
C110	27412	27462	+				TTTTTCCCCAGTTCAAATCCGGGTGCCGCCTCAGCAACAAACTTTAAATA
C111	31549	31569	+				TAATAATAAAGTTAGCGAGT
C112	35969	36002	+				GACCCCCTCCCATTCCTTGAATTACACATTCAA
C113	42029	42048	+				TCTGGGGCAAGTGTTCGGA

Name	Start	End	+/-	Start	End	+/-	Sequence
C114	43053	43072	+				TAGAATTTTCTGAAAGGTA
C115	44676	44702	+				AGTACGAACTAACATAAAAGCGGACT
C116	45429	45456	+				TCGGGGTTTGCAGCGATAACTTGGTAT
C117	46627	46645	+				AGAATCGACCGTTCGACT
C118	47446	47493	+				CGTTGACTTTTAAAATCGTGAGGGTTCAAGTCCCTCTATCCCCAGCT
C119	52102	52131	+				CATTGGTTCAAATCCAATAGTAGGTATAA
C120	56456	56488	+				TTTTTTTTACTAAAAAAGATTGAGCCGAGGTT
C121	56489	56523	+				TCTGTTGTATATACTATTTTTTTTGATAGATACA
C122	59448	59475	+				ACATAGATTCCTACAAACATAAATAAA
C123	61698	61721	+				ACTCGCTCCATATCTGTCTCACT
C124	62162	62188	+				CCCTGCTACTAATAAAGATGTTCACT
C125	63038	63074	+				AAAAGGAATTTTAGACATCCTTTTCTTGTGTCGATC
C126	66334	66349	+				GGCTAGAAAGAGGGC
C127	66925	66944	+				CTAATGCGAGATCTAAAAA
C128	67093	67119	+				AGAGATACAATCAACAATCGGGGACT
C129	76317	76358	+				CTTACTTATTACTTGGTGAAGGAACGATAGTATTTTATTGC
C130	79773	79790	+				ACTTATACTTAAGAACT
C131	87931	87955	+	150693	150717	-	GGGAATCCTACAAGAGCCATTCGT
C132	93735	93750	+	144898	144913	-	AGATAGACCTTTCTC
C133	95045	95067	+	143581	143603	-	ATTGTTTGATCTTAAAGGGGAT
C134	95437	95459	+	143189	143211	-	ATCCACCATTTGAGTCTCCAAC
C135	100407	100441	+	138207	138241	-	CCTACCCGTATCGAATAGAACATGCCGAGTCAAA
C136	100759	100779	+	137869	137889	-	GTTCGAGCCTGATTATCCCT
C137	102981	103008	+	135640	135667	-	ACTTCTCCTCAGGAGGATAGATGGGGC
C138	106393	106441	+	132207	132255	-	GAACTCGGCAAAATAGCCCCGTAACTTCGGGAGAAGGGGTGCCTCCTC
C139	107481	107500	+	131148	131167	-	CCAAGATGAGTGCTCTCCT
C140	107662	107701	+	130947	130986	-	ATGCAGCTGAGGCATCCTAACAGACCGGTAGACTTGAAC
C141	108035	108070	+	130578	130613	-	GAGGTCCTGCGGAAAAATAGCTCGACGCCAGGATG
C142	108170	108197	+	130451	130478	-	ATCCCACTTCACACCCCGGAACGCACC
C143	109228	109257	+	129391	129420	-	TCTATTTCATTATATTCCATCCATATCCC
C144	109330	109351	+	129297	129318	-	TCTATATATGGAAAGTTGCAA
C145	114327	114352	+				GGTTCGAGTCCGAGTAGCGGCATAA
C146	114692	114710	+				TTTTTCTTTCGTGGGCTT
C147	121695	121715	+				GCTATAGATGGTCCAATACT
C148	123601	123623	+				ATTAATTTTACTGATCAGTAAT
C149	127470	127492	+				TTATAAGCGTTTGATCGTTGCT
C150	129927	129974	+	108674	108721	-	ATTACCGCGAGCAAACATATGAATTTAATGACTTAATGATGAGGAAC
C151	130051	130068	+	108580	108597	-	AAATATGCTGATTCGGC
C152	144032	144062	+	94586	94616	-	TTGGGACCCTATTCACCTCTTTGGTTGGAC
C153	144332	144376	+	94272	94316	-	TCGTGCTAAAGAGCGTGGAGGTTCGAGTCCTCTTCAAGGCATAA
C154	42	77	-				GGCGGATGTAGCCAAGTGGATTAAGGCAGTGGATT
C155	1722	1737	-				GTTCGAGTCCCGGGC
C156	3350	3374	-				TTTTGCGAATGTATGAACAGAATC
C157	4316	4347	-				GGGTTGCTAACTCAACGGTAGAGTACTCGGC
C158	4364	4393	-				CTGATTGTATCTACATATTTGCAGTACGT
C159	6656	6687	-				TGGGGCGTAGCCAAGCGGTAAGGCAACGGGT
C160	7851	7872	-				GGAGAGATGGCTGAGTGGACT
C161	12966	12990	-				CATTATTATTATTGAAAATTAAAA
C162	13526	13552	-				ATTGTATCATTAACTATTTCTTTATT
C163	14798	14818	-				GTCTTGAATCAAAATAATTT
C164	20095	20115	-				ACTCAAACTCATTGTCGAAT
C165	22319	22350	-				TTTATATAAAGTAAACAAATATGTCATGGTT
C166	22688	22718	-				CCTAGTTATATTGCGAATCTTTTAGATAAA
C167	23155	23190	-				AACTACGATCTTTGGCTCTGGAACTGAATCATTTC
C168	26441	26466	-				TAAAATTTCATGTGATTCAGTAAAC
C169	27508	27529	-				AGAACCTCGCGAGCCAGGGGC
C170	28566	28585	+	28587	28606	-	GCTAGTATGGTAGAAAGAG

Name	Start	End	+/-	Start	End	+/-	Sequence
C171	28960	28988	-				GACGATGAATCGATTTTATAGCTCCGAT
C172	29841	29874	-				GGGATTGTAGTTCAATTGGTCAGAGCACCGCCC
C173	30383	30406	-				GGGTCGATGCCCGAGCGGTTAAT
C174	30520	30538	-				GCCCCCATCGTCTAGTGG
C175	30520	30563	-				AGTATGATGGCGGTTGAGCAAGTATGCCCCCATCGTCTAGTGG
C176	31562	31591	-				GTTCATTTTTATTTTTTAAATTACTCGCT
C177	35267	35311	-				TTTGTTGGATGAATCTATTTTTCTCTTTATTGGCTTTTTTTACT
C178	35368	35403	-				GGAGAGATGGCCGAGTGGTTGATGGCTCCGGTCTT
C179	36760	36777	-				CGCGGGGTAGAGCAGTT
C180	37273	37306	-				ATTTATTTCTACATCTAGGATCCGATTTGTATC
C181	44673	44700	-				TCCGCTTTTATGTTAGTTCGTACTATA
C182	45300	45322	-				CCAATATGAAGGGTTAGTCAAT
C183	46747	46780	-				GAGCTCCATCCATATTGATTTAGTCTAATCAAC
C184	47136	47155	-				GAAGTTTCGATCGAAGGAT
C185	49633	49657	-				ATGAACAAATGCCTGAACCGAAGT
C186	50692	50715	-				ACTGGATTTTTTGATACGTCATC
C187	51842	51871	-				AGGGCTATAGCTCAGTTAGGTAGAGCACC
C188	54212	54240	-				TTGAATTAACCGATTAATTTGCTATCGA
C189	54314	54336	-				TGAAAATGACTATTCCTTCATT
C190	57585	57609	-				AATTATTACTATCGATTAAAAAGT
C191	57711	57740	-				TTGAGTTATATCGAAATCCTTAGAACTTA
C192	58153	58182	-				GCATTCGTGCTCCTCCGGAAGAACACACT
C193	58401	58418	-				TGAACCTTCAGGCACGG
C194	61145	61169	-				ATATAGGAATTCTTGAACCCAAGA
C195	63190	63211	-				TCTGATTTTATTTATTTAGTA
C196	64418	64447	-				GTAGACTCTAAAAATACCCTTGGTACTTT
C197	66229	66253	-				GTAGGTTCAAATCCTACAGAGCGT
C198	66266	66302	-				ACGCTCTTAGTTCAGTTCGGTAGAACGTGGGTCTCC
C199	66546	66563	-				AGGGATGTAGCGCAGCT
C200	67562	67583	-				TCTGGAAATTCCCGCGGCTTC
C201	68138	68160	-				TCTGATTATTAAGAAAAGGTAA
C202	68357	68379	-				ACGAGTTATGCTTTTCGACGAT
C203	69743	69772	-				AGCCGGTTAGAACTAATCTAAACCAGCCC
C204	71915	71940	-				TTTTACGTTTCCACATCAAAGTGAA
C205	76598	76614	-				TGAACCAGCCTATCCC
C206	76940	76961	-				TAAGTGCTTTCTGGGTCGTCT
C207	76987	77009	-				ACAGGTAAATGCTCAACACCCA
C208	77146	77167	-				CCCCCGAGGGAACCGGACATG
C209	79289	79309	-				GTAGAATACCAAAGGGAGTT
C210	82670	82712	-				ATAAGCAATTCTATAAGATTGAATAAAAATTTCCATCAAAAC
C211	112046	112070	-				TATTAGGAATTTTAGGTCTTTATT
C212	113825	113843	+	113802	113820	-	GAGGAAATAAAAGATCTT
C213	114638	114657	-				CCTGAATAAATCCAACGAG
C214	117698	117725	-				AGAAATCAAAGTATTTTAGCCCCATTT
C215	118935	118972	-				AATTTCCTGGTTAAATTAATAAGGTCATGAAAAGGAT
C216	120515	120543	-				TCGGGACCCCAGATATATTTAATCCATT
C217	123606	123630	-				CATTATTATTACTGATCAGTAAAA
C218	126503	126526	-				TTTGAACCTATTTCTAAAGAATT
C219	2954	2977	-				AAAATCAATTTTGAATCCAAGAT
C220	6224	6245	-				AACTATGACTATTCATGATTC
C221	7588	7607	-				ATACAAAAAGTTTGAGAGT
C222	7781	7817	-				ATCGTACCGAGGGTTCGAATCCCTCTCTTTCCCCTT
C223	9587	9605	-				ATTAGACAATGGACGCTT
C224	11488	11535	-				TGATTAATTATTTCCTTACGATTATTATAGGCATTATTTTTTTTTCT
C225	28564	28606	-				GCTAGTATGGTAGAAAGAGATCTCTTTCTACCATACTAGCCA
C226	35214	35234	-				CTGGATAGTATAGCCGAGCC
C227	36827	36859	-				TTCTATTTGTACAGATATGGAAGAGGGGCTCC

Name	Start	End	+/-	Start	End	+/-	Sequence
C228	37248	37279	-				TGTATCATTATCATTGATAATAACAGGAACT
C229	42409	42436	-				AATTTCATTATATCCTTTTCTCAAATC
C230	46140	46158	-				AAGAAGATTGAAAAGACT
C231	46212	46234	-				GGTTCGATTCCGATAGCCGGCT
C232	48599	48638	-				AGAGTTCTGCATTATGAACTTTGTATCGCGCACATAACT
C233	50561	50585	-				TATTAGTAATAGAAACATGGAACT
C234	60550	60582	-				TTTGAATCTAGAAAGAATACAGAAACAGACTC
C235	63442	63468	-				TCTGATTCGAGGGGGTCCCGTTGAAC
C236	66486	66523	-				CAAAATGTCACGGGTTCAAATCCTGTCATCCCTACCT
C237	68090	68110	-				GCAATAGTGATTAATCGTTG
C238	70805	70822	-				TGTTTATAAACTCTCCT
C239	77716	77736	-				TGACTACTCCCTAGATACCT
C240	82643	82663	-				ATATAATTGCTATGCTTAGT
C241	114222	114246	-				AAGTTTTGATTCAATCGTCGAGAT
C242	117596	117614	-				TAGAAGTTTACTAGATTG
C243	119787	119813	-				TTTAAACAAGAGACAGAAACAAAGAT
C244	126715	126745	-				ACAGAATTTCCAAGAAACTGGTTAACGGAT

Supplementary Table 3: Small RNAs identified in mitochondria

Name	Start	End	+/-	Start	End	+/-	Start	End	+/-	Sequence
M1	15281	15305	+							GTGGGAACTCTACTTGCCATTCCT
M2	16888	16910	+	280184	280206	-				CCTGAGCTCATCAATGAACGGA
M3	17033	17057	+	280037	280061	-				TCTGGATCCCCGAGAGTTACTCCA
M4	22511	22527	+							GTATGGAAAGACGCCT
M5	41347	41363	+							GCTGGAATAACTCAGA
M6	44371	44387	+							TTTTGAAGGCCTTGGC
M7	46915	46940	+	181080	181105	+				GAAGAAAGATCGTTTTTAGATCATC
M8	46940	46961	+	181105	181126	+				AAGTGAGGACAGGTAGTAGCT
M9	52147	52181	+							GAAGTAGTCGTCGTCTGACCAATTGACTCGGACA
M10	62348	62364	+							GGAGAGATGGCCGAGT
M11	71355	71396	+							GAAGAAGGTTGACAAGAAGAATAATTTGTCTCCTGTGATTG
M12	71452	71485	+							AGCGGGGTAGAGGAATTGGTCAACTCATCAGGC
M13	81385	81400	+	37339	37354	-				CCAGCAGCCAAACCA
M14	92441	92470	+							AAGTAAATAGTCGTCAACTATCGAGAACC
M15	98945	98977	+	12276	12308	-				GAAGAAAGATCGTTTTTAGAAAAGAAAGAACG
M16	99028	99059	+	12194	12225	-				AAGTGGTAACAGGTAGTAGCTCTGGTAGAGT
M17	99107	99139	+	12114	12146	-				TAGTTAGTTTCATCGATATTTTTGTGGTGTTC
M18	103758	103801	+							GAAGAAGATTTTAATTCCAGCTTAAATAAGTAAGACTTGACTC
M19	103826	103852	+	227108	227134	+				GGAGGGATGGCTGAGTGGCTTAAGGC
M20	104164	104186	+	227446	227468	+				TAGTCAAGTGGTAAGGTAGGGC
M21	104220	104254	+	227502	227536	+				AAGTGGTTCAGCTCAGCTGGTTAGAGCAAAGGAC
M22	104295	104333	+	227577	227615	+				TATTCTCGGAGCTGAGGTATATGAAGAATGGCCTTTTG
M23	104456	104481	+							CGAGGTGTAGCGCAGTCTGGTCAGC
M24	104884	104916	+							GGCTAGGTAACATAATGGAAATGTATCGGACT
M25	105088	105111	+							TCTGGCTAACATTGAACTTGGTA
M26	105581	105610	+							TCTGGATTATTATATGATGATTTTGCAAC
M27	105727	105742	+							AGAGCCGCTTTGAGG
M28	105831	105854	+							GTTGAGAACGGGAATTGAACTCT
M29	105876	105908	+							TCCTCAGTAGCTCAGTGGTAGAGCGGTCGGCT
M30	106796	106837	+							GGGAGAGTGGCCGAGTGGTCAAAAGCGGCAGACTGTAAATC
M31	106954	106972	+							GGAGGCAGGCTTGGGGGT
M32	107061	107078	+	279314	279331	+	143130	143147	-	TCAAGCAAGTTGGGGAA
M33	109542	109560	+							GTCTCGGTAGGACTTCCA
M34	111593	111611	+	129751	129769	+	298179	298197	-	GAAGAAATCTCTATGCCC
M35	120168	120185	+							TGCAGCCCAGCTGGATC
M36	121453	121475	+							GAAGAAGACCGGTTAGGATCAC

Name	Start	End	+/-	Start	End	+/-	Start	End	+/-	Sequence
M37	130456	130506	+	170980	171030	+				GGGCGAGGATACTTGCCTTCGCGGTTCGACTTTCTTTTCAGGCTTGACTC
M38	133224	133248	+							ACCTTATTCTGATCGTTCAGAGGG
M39	137712	137736	+							TCTGAACGAATTAGATCCTTGGTA
M40	138161	138205	+							GTTGCTCGATCAGGACCTTAGCTTTATTGCGAGCCCAGAAGTCT
M41	141598	141616	+							TTTGAAAGAGAGTAAAAC
M42	146559	146583	+							TCCATCCCTGAATGTCGTCGGGTA
M43	146802	146826	+							AGATAGGTCGTCTGAGGGTTCGCC
M44	155132	155153	+							CGCTTATGTAGTGGTCGGCCT
M45	155982	156006	+							TGAACATTTCTAGTCACACGGGAA
M46	158319	158342	+							CCCCTAGAACCTGGCAAAGTAAC
M47	130484	130506	+	170768	170790	+	171008	171030	+	ACTTTCTTTTCAGGCTTGACTC
M48	187504	187527	+							GTCACACAGGTCGTCTTTGGCTC
M49	187740	187764	+							AGGGAAAAACTGCCTTGGAGGCTG
M50	190554	190578	+							TCTGGAAGCCCGGCTCGCGAGGCG
M51	190668	190692	+							GCGTCAGTTTGTGGATCGCGGTCT
M52	191383	191407	+	261553	261577	+				GGGGCACTTGATTTACCGAGGGTT
M53	194958	194976	+							ATCCAATAGTAGGTAACT
M54	210984	211008	+							AAAGAACTCAATGAAAAAGGGCCT
M55	212671	212695	+							AGAATGGAACCCTACTCTGAGAGG
M56	217637	217669	+							GAAGAAATCAAGTTGATAGATCAGTTAGTTGA
M57	217832	217856	+							AATCCATCTCGGTCGAAGAGCTGA
M58	219109	219133	+							TCGAAGACAAAGAGAACCGGGCTT
M59	219960	219984	+							TAACCAGCGTCAGGTCGAACGAGC
M60	221768	221792	+							GTTATTTCCAGGAAAGTTGAGATC
M61	222825	222849	+							AAAAGATAGTTCCGATCGTTGAGT
M62	223826	223850	+							TCTGCGAAGATAGAAGAGCGGACT
M63	227615	227643	+							GTCCCTTTCGTCCAGTGGTTAGGACATC
M64	227687	227704	+							GGTACTCATTCTCGGCC
M65	234579	234603	+							ACCGGAAACCGTTTGATCAGGATA
M66	235627	235651	+							ACCCAATTCCCGTGATCGAGGAAC
M67	242996	243020	+							AGATATAGATCGGTTGGCACTGGA
M68	255163	255180	+							GTGAGGTGGGGAAAGGT
M69	261337	261361	+							TGATCTCGAATAGATCTTCGGCCT
M70	263112	263136	+							AGCATTGCCACTTGCTTCAAGCTG
M71	267351	267375	+							ACGAGAGAGTGAGATTAGACTGCT
M72	273992	274015	+							TCTATATTCCGGGTCCAAGGATG
M73	275068	275118	+							GAAGAAGGTTGACAAGAAGAATAATTTTAAAACTGGGATTGTAGTTCAAT
M74	275102	275135	+							GGGATTGTAGTTCAATCGGTCAGAGCACCGCCC
M75	275963	275987	+							AAACCCAGAAAGGTCGTATCGGTC
M76	276948	276972	+							GACATATCACAGTAAGTCGATAGT
M77	277040	277064	+							GGCGAACTGGAACATATGTCGGCT
M78	277441	277465	+							AGAAGAATTAGCTGATGTAGAAGG
M79	278655	278700	+							GAAGACGAAGACGGATCAAATTGAATAATCGAAGAGAGATGGGAC
M80	278811	278849	+	17836	17874	-				GAATGCATTCCAAGTGAGATGTCCAAGATCAAAGGAAC
M81	280227	280251	+							AATGCCCGGCATTACGTCGACTGA
M82	280671	280695	+							GTTGGTAGGCTCCGGAGAATAGAA
M83	282011	282035	+							GCCTGGACTGAAAGGATTCTCTTT
M84	282475	282497	+							CTGTCCGGGATCTCTTCACTGA
M85	282551	282575	+							AGCACATGGACCGGATTGTTACTC
M86	289391	289415	+							ATCTGAGGAGGAGGCTTCGTCGTC
M87	291642	291666	+	118060	118084	-				TCTGATTGAGTGAACATACCGAGT
M88	301172	301196	+							TCTACAGGAGAAGTCGCTTATGGA
M89	309149	309173	+							AGTAGCAAACTTGATTCTGTGGCT
M90	312406	312430	+							AAGTATGATTTGATTCTAGGGCCT
M91	314574	314598	+							ATCGGCCTCGTCATCGAAAGCGGC
M92	317460	317484	+							AAGGATAACTGTAGGTCGGTGGCT
M93	319112	319127	+							GGCTGTGTTGGGGAA
M94	321926	321950	+							GATATACGACATCGTTGGATCCGA
M95	323139	323177	+							TCGCTAGAGCTGAAGAAGTTTCGGGCTGAAAAGCTGCC
M96	328657	328681	+							GTAGAGAATGAAGAGGGGCCTAGG

Name	Start	End	+/-	Start	End	+/-	Start	End	+/-	Sequence
M97	329393	329409	+							TTTGGGGCCCTTCATC
M98	329623	329647	+							ATGCATTCCCATCGGTGTCGCAAC
M99	329665	329688	+							GTCCGGTGCACGGAGAACTGCCT
M100	330451	330468	+							TGAACCGAACGTGAAAG
M101	334687	334734	+							AACCAATCAGTAGTTGATAGTCAAGGGCGTGTTATTAATACTTGGGC
M102	334776	334809	+							TGAACGTAATGCTCACAACTTCCCTCTAGACCT
M103	337668	337700	+							ACCTACTTGACTCAGCGGTTAGAGTATCGCTT
M104	337719	337742	+							GGTTCAAATCCAATAGTAGGTAA
M105	340296	340320	+							AACACTATCGGTAGTCAAAGGAAG
M106	340997	341020	+							GTAGTATTCCATGAGTTGGGCTA
M107	346454	346470	+							CGGAGAAGGGGCTCCA
M108	278853	278897	+	346715	346759	+	17788	17832	-	GTAAGAATCGACGAGGAATCAATAAGATATAAGATAAGTGAATG
M109	365082	365126	+							GCGGTACCAAATCGAGGCAAACTCTGAATACTAGATATGACCTC
M110	365144	365186	+							GTCGGCCAGTGAGACGGTGGGGGATAAGCTTCATCGTCGAGA
M111	16945	16966	+	280128	280149	-				CCTGAGCACAGTGAAATGCCT
M112	19875	19899	+							CTTCCCTGATGATCTTGTCGGCCC
M113	31142	31163	+							ATGTAAGCCATGTATCTAGGA
M114	41406	41430	+							AGAGGAACAGTACGATCTTGGACT
M115	46987	47012	+	181152	181177	+				TTCTGTCTGCGGTTCGAATCCGGAC
M116	60920	60956	+							TTATTTTTATGTCAAGGATCTAGTTGGTTGGGTAGC
M117	61428	61475	+							TCTGACACCAATCATTTACATATTACACCAAGAATTGACAAGCAGAT
M118	71505	71529	+							GTTCGAATCCTGTCCCCGCATAAA
M119	103787	103826	+							GTAAGACTTGACTCTTTAAAAAATTCCGATCAACAACTT
M120	103877	103917	+	227159	227199	+				GAAGATTGTATCATGGGTTCGAATCCCATTTCCTCCGGCG
M121	104510	104534	+							GTTCGAATCCTGTCACCTTGATTA
M122	104934	104958	+							GTTCGACTCCGTCCTTGGCCTACA
M123	106859	106882	+	333647	333670	-				TTCGAATCCTGCCTCTCCCACTT
M124	112952	112978	+	296748	296774	-				GGATGGATGTCTGAGCGGTTGAAAGA
M125	113018	113042	+	296684	296708	-				GTTCGAATCCCTCTCCATCCGCGA
M126	113136	113176	+	296550	296590	-				TAGGAACTTTGTCTCCCTTTCGTTATCTTCTCTTTTTTTC
M127	137433	137457	+							GTTGGCTTAACGAGCGCAGATGTG
M128	146885	146907	+							CCTACCACTAGTCTTCGGCCGG
M129	153843	153867	+	263870	263894	-				GAAGCGAGATCGGAGTAGGAAGAC
M130	155170	155194	+							CTTTGTCTTCGTCTAAGAGCGCCT
M131	157410	157452	+							CTTGAGATAAATTATCAAATAGGAAATTGCATACCATTAGCC
M132	168421	168445	+							AACAGAACAGAACCACCGTAAGGA
M133	169686	169705	+							AAGAAAAGAAAAACGGGTC
M134	172342	172364	+							TGCAGAGATTCGGATAAAGCTC
M135	194896	194927	+							CCTACTTAACTCAGTGGTTAGAGTATTGCTT
M136	198414	198436	+							AAGCGAGAAAGGGGATTGGCTC
M137	222283	222307	+							AAAAGAACTCCCTTGAGCTTGGTA
M138	222363	222387	+							GTAGCCCCTCTAGCTTGGAACCCT
M139	224496	224520	+							AGACAGTAGGCTTCCGGTAGGGAC
M140	227807	227831	+							AGAAGAACGAGACACTGTAGGCTG
M141	254338	254362	+							ATTATTGCAGTAGGATAATGGCTA
M142	261311	261335	+							AGCCATCCGTTGGATGATTTGGGC
M143	261712	261734	+							GGAAGAACTGCCCGACCCGGAA
M144	275068	275102	+							GAAGAAGGTTGACAAGAAGAATAATTTTAAAACT
M145	275144	275179	+							GGAAGCTGCGGGTTCGAGCCCCGTCAGTCCCGACC
M146	279314	279332	+	143129	143147	-				TCAAGCAAGTTGGGGAAC
M147	279960	279982	+	17112	17134	-				GTAGAGACTATCACGAGCGCCT
M148	280285	280302	+	361378	361395	-				GTCGTAACAAGGTAGCC
M149	314442	314466	+							TCTGGAACTCAGGGAGCAAGACCC
M150	315392	315424	+	270407	270439	-				TCTGATAAATGCACTTCAAAGGGAGGGAAGGC
M151	328756	328780	+							GATAAGGAATAAGGATTGAAGCCC
M152	331964	331986	+							CGGAAACTCGAGAAGGTCGCCT
M153	332374	332398	+							ACGGTACTAAAGGTCCTCGGACTT
M154	334610	334640	+							CTTGGCCGGTAGTAGGTATTTGGTTTACTG
M155	334687	334713	+							AACCAATCAGTAGTTGATAGTCAAGG
M156	340761	340783	+							CAGAATGAAGGTCGTCAGTCCC

Name	Start	End	+/-	Start	End	+/-	Start	End	+/-	Sequence
M157	340876	340900	+							GCCCTTGCTGTTTCTTCGACTGTT
M158	363063	363087	+							AGGATTTGGCCCGAACTGTTCGGC
M159	12069	12087	-							GAATGCATTGGATGGATG
M160	15758	15782	-							CAGATCGGATCCAATCAAAGCTGT
M161	16737	16777	-							GGGCCAGACCGGTGTCCATCCCGACCACGGGATGTTGGGC
M162	20631	20658	-							GTGGGACTCTCTATCTTCTTGGGTAGA
M163	22840	22870	-							AAGAAAGAATTGACAAGCGCATAAGTTTTC
M164	27637	27652	-							GTTGCGAGGGCCTTG
M165	28930	28970	-	204372	204412	-				GGGTGTATAGCTCAGTTGGTAGAGCATTGGGCTTTTAACC
M166	38500	38515	-							CTTCGGTCCGGGGGT
M167	42728	42776	-							GAAGAATCTTACGCCCCAAATTCCCATCTCTTTTTTCTTGGTTGGACC
M168	45291	45310	-	179456	179475	-				TCACAGAGTCATCGGTATC
M169	51180	51208	-							GCTTTCAATCAATAGAAATCGTATTCGT
M170	53699	53735	-							CCTGGTGTTCGAACTAGTCATTAATGGTCGGCTTCA
M171	53776	53807	-							GCGGAAATAGCTTAATGGTAGAGCATAGCCT
M172	70621	70638	-							TCCTCTTGGAAGCACCA
M173	72038	72053	-							ATTCAAGGAAGCGGA
M174	77397	77438	-							GAATGCATTAAATGGATGCATTGAGATTCCGTAAGTAACTC
M175	80903	80927	-							TCTGGAAGGCACATGAGTCCGAAC
M176	83083	83115	-							TGTGGCTGCTTAAAAAACTGATTCAACGAGAT
M177	86631	86655	-							AGGCACTGCCGGAAACGGGACTGC
M178	132736	132760	-							GCCTGCGGCGTTTTCGCCAGACGG
M179	135810	135825	-							GGGATGGGTAGGCCC
M180	144228	144254	-							TTACTTATGAGATTAGTTGAGTAGAC
M181	144839	144863	-							CGGAGCAACGCGTCATCGGACGTT
M182	150863	150887	-							TCTGAGGGCCTTTGTTTTGATGAA
M183	152923	152947	-							ACATCAGTGATCGGCAAACACAGG
M184	152999	153023	-							ACTGGAATACAATGAGACGTTGAT
M185	154129	154144	-							TGCCGGTCGATAAGC
M186	154991	155015	-							GGAGCGAACTTCTCGATCGTGTTC
M187	155149	155171	-							GGAGGCATTCCGGTAGGAAGGC
M188	162340	162355	-							TTTGTCAGGTTCGGT
M189	176313	176352	-							GAGAAGAACGTATCAGCAACTCGACGAAAAAATGGTAAA
M190	186508	186531	-							GTAGAATCAATCAACGGCACCTA
M191	187176	187193	-							TAAATGGTTTTGGCGGA
M192	191923	191953	-							TATTGTAAGCATTCCTCGGAAGAGCTCGCC
M193	191991	192025	-							AAGTGGTTCAGCTCAGTTGGTTAGAGCAAAGGAC
M194	197959	197983	-							AGGAAAGTTCCTCAGTCAACGAAC
M195	203709	203735	-							TAATATCTGGCGTCGTCAGGCGTTGA
M196	205862	205896	-							TTTTGATGGAAGAACAGGAGATCCTTTTGAACAG
M197	210205	210229	-							AACTAAGATTCCATTCGTCGAAAC
M198	210761	210785	-							TTTCCGCTTTGATAGATAGATCTG
M199	211515	211539	-							CAGCAAATCGAGGTCTCGACGAGC
M200	211584	211606	-							TCTGAGCTTGGTGTTACGTGGA
M201	212439	212461	-							GTTTCAGAATTCCCAAGCGCCT
M202	213118	213142	-							TCTGACTATTACCCGGGAACGGAC
M203	213305	213329	-							GAAGGAGAAAAGGATGGTGAATTC
M204	213574	213597	-							GTCGGGAGTGAGCCTAGCTTCCC
M205	215926	215950	-							TCTGATTTCTCATATTACCCGGGG
M206	217821	217845	-							ACCGAGATGGATTTGTCGTTGGAG
M207	218963	218987	-							AACGTGATGCTCCTTCGTCAGATG
M208	219394	219417	-							GTTCCGATATCTTTCGTAGGATG
M209	219945	219969	-							CGCTGGTTAGACGTGAGGTCGAAC
M210	220198	220222	-							AGTCCAAGACTCTTTTAGTAAGAC
M211	220452	220474	-							GTAGAATCCATCTAAGTAGCCT
M212	221439	221462	-							GCCCATAGCGCATCGTCAAGCTT
M213	221766	221783	-							TTTCCTGGAAATAACTT
M214	222770	222794	-							CTGAACTTGGGCGAGAGATGTGAC
M215	223626	223650	-							AGTGAGAATACTGAACAGACAACA
M216	231292	231322	-							GTCACGATCAGTCTAAGGTTGAAATCTGGA

Name	Start	End	+/-	Start	End	+/-	Start	End	+/-	Sequence
M217	232463	232487	-							AAAAATCCTCTGGACGCTTGGCGC
M218	234387	234402	-							TGAAGGTCGATCATC
M219	236938	236962	-							GTTGGATACCCACAGTCAGAAGAC
M220	237139	237163	-							AGGAAATCCCTTCTGAGTTGGACC
M221	240008	240023	-							GCATCCATGGCTGAA
M222	240033	240080	-							ACGGAAGAAATTGAAGCTCGAGAAGGAATACCAAAACCTAGTTCACT
M223	240421	240436	-							GGGGGTTTCGGGGAA
M224	247870	247914	-							GTATATTCTGGGCGAGGAGCGTAAGCGACATGGCATATTTGTGA
M225	249943	249959	-							TCTGACCAGTGGTGCT
M226	250080	250104	-							GTAGGTTCAAATCCTACAGAGCGT
M227	254466	254503	-							TTATGAACACCCGATCGGATCTGTCAAGAACGAGCTG
M228	254502	254538	-							CCGGCATGCAAAGGTTCGAATCCTTTTACTCCAGAT
M229	255014	255034	-							CCTTACAAAGGGAAACGGCC
M230	261799	261823	-							ATCCACCTAGTGGGGGGTCTGGCT
M231	262112	262129	-							TATGCGTTCCTCGGACG
M232	262210	262240	-							GAAAGAGATTCGTTGGATAAGTTGAGAACA
M233	274734	274755	-							TAAGCTAGAACTGCTCCTTCT
M234	276673	276696	-							GACCAATTACGATCGATTCGCTA
M235	277813	277837	-							AAGCACTCAACTTGATTGGAGAAG
M236	280261	280285	-							TTCACCCCAGTCGAAGATCCCAGC
M237	288223	288247	-							ACAATGCTCTGAACACGAGAGTGT
M238	288595	288619	-							TCTGAACTGCGAGAATAACTGACT
M239	288806	288830	-							TCTGATCAAGGGCCGGGGCACACG
M240	289626	289646	-	363213	363233	-				ATGCTTAACACATGCAAGTC
M241	303209	303250	-							CTATGCAAACAGGGTTAAAAGCGGTAGATAGCCTGGTTCCT
M242	306507	306531	-							TTCCAACCCCTTGGAAGAGAGGAA
M243	306943	306967	-							GTCGAGAACTCTTAGAATTGTGCT
M244	311640	311664	-							GTTTCGGATGATGAATAGTCACTC
M245	315630	315654	-							CAACGTAGGTCGGTAACAGATTTG
M246	316356	316380	-							GTGCCGAGCATTGTTCGTCGTGCT
M247	316700	316717	-							GCTATGGACTTAAAAGC
M248	326458	326482	-							GCTGTTGGTACAACTGTCATCGGT
M249	326552	326576	-							GCTAAAGATCAGTTTCGGTTCTAG
M250	329690	329714	-							ACGAATAAGTAAGTTTGGAGGACC
M251	330430	330454	-							TCAGAGAGCACTTTTTTCGTTGAG
M252	332600	332622	-							GTAGAATCACGCAACGCACGCT
M253	334435	334464	-							CTGATCAAACTAGAAGTTACCAAGGAACC
M254	340124	340139	-							GCTCTCTTTCCGCCA
M255	341416	341440	-							ATGACGAAAACAATGAGGCGGATT
M256	341661	341685	-							GACTTGAGATTATTGGATTGTGCC
M257	342302	342326	-							CTGCTATGCTGAGAAGTCGGCTGG
M258	349782	349827	-							GAAGAAAAGGTCGCCGACTGCTACTAAGAACCTAACAGAACTTTT
M259	351083	351100	-							TCTGGTTGTTGCCACCA
M260	351728	351764	-							ATAATTATGTTCTGTCGAAATGTGTTTGTTGTATTT
M261	354178	354202	-							AGGGAATTCCTAAGATCAGAACTG
M262	359244	359280	-							AAGAAGATTCGAATTCCAGTCACTTTAGATATCAAT
M263	359662	359686	-							GTTCAATTCCCGTCGTTCGCCCAT
M264	359720	359740	-							GGCGGATGTAGCCAAGTGGA
M265	359872	359894	-							GAAGTGGAGTGGTGAGGCGGGC
M266	361154	361180	-							CAAACCGGGCACTACGGTGAGACGTG
M267	361528	361545	-							TGTACACACCGCCCGTC
M268	362300	362319	-							GCAAAACCTTACCAGCCCT
M269	362340	362359	-							GCACAAGCGGTGGAGCATG
M270	363255	363283	-							TCATAGTCAAAAGAAGAGTTTGATCCTG
M271	363309	363353	-							GAAGAAAGGTCCACAGAAGGTTGGGAAGTAGTACGCCCGGTTCA
M272	10480	10502	-							ATTGGATGATCGGGCCGAGGGC
M273	16719	16746	-							ATGTTGGGCTTCAACTTCCCTTTGGCC
M274	122544	122561	+	18183	18200	-				AGAAATGATGGTTGACT
M275	20484	20515	-							CATTCCCAGTTCTTTCTCTTCTCTCTTTTTT
M276	20550	20598	-							TTTTTATACAAAGTCAAGTCAAGAATAATAATCGAACTGGAGGAGCTT

Name	Start	End	+/-	Start	End	+/-	Start	End	+/-	Sequence
M277	20735	20754	-							ACGTCCGGTTCGGAGGGCG
M278	23607	23641	-							CCAGTCCAGGGGACAAATCAATAGGAAATGCTAT
M279	28581	28626	-							TTTTCCTTCAGTTTATCCTATATTTTCAAAAAAGCGTGGGAGGAC
M280	28906	28941	-	204348	204383	-				GGCTTTTAACCTAATGGTCGCAGGTTCAAGTCCTG
M281	76657	76683	-							ACTATAATGAGGAGGACGACTGACCC
M282	81529	81557	-							CTTATGTCAAAAGGACCAAGGACGATCT
M283	105454	105475	-							CTCTATTATGGATTTCTGACC
M284	127017	127051	-							TATATTGTAGGTTCGAGCCCTACTAAGCCTACCA
M285	140029	140050	-							CTATAACTCTGGGAACCGGGG
M286	143236	143282	-							TCTATCATTAGCTCGGGTAGTCCCTGTTTCTGGTCTTTTAGTCACC
M287	145691	145715	-							TCCCTTGTTTCGTCGTGGCACACC
M288	147867	147884	-							GAATAGATCCGTGGGCC
M289	155340	155364	-							TCGCGGAGCGAAGAAAGCGGGCTT
M290	168807	168834	-							CCTGTTGTTCTGTTCCTGCCACGAGAA
M291	188011	188047	-							TGCTCTCAGAAGAGCGGATCCAATACCAAGACTACT
M292	188772	188814	-							GGAAAATACGAAGTTCTCTTCTCCTTTCGTTCTCTTTTTTTC
M293	197064	197088	-							TCGGGGATTCGGATGTTGAGATGC
M294	219482	219505	-							GTGGGATTCTGAAACGTATAATT
M295	230693	230714	-							AACCAATGGAGTTGATTACGT
M296	237911	237939	-							GTTGAACGAGAACTTTATAATTAAGCCT
M297	242028	242059	-	260676	260707	-				ATTTATTTTGACGATTGGATTTCTATATGAA
M298	256694	256734	-							TCAGTAGATTATTTAGAACTTCGGAAGATGGTCAAGGTAC
M299	279882	279906	-							CTCCTCAGGAATCGGTTGATTGAC
M300	280014	280061	-							TCTGGATCCCCGAGAGTTACTCCACGTTGATGCAAGAGAATTTGGGC
M301	286519	286554	-							GTCGAGAGAGTACGATACATCGGTGTAAAAGGTTG
M302	290854	290878	-							AGAGACGGTTGACCGAGCGGAGAC
M303	317073	317097	-							ATCGGCATACTCAAAAGGAGGCGC
M304	327853	327889	-							GTTCGATCATTGACAAGGTTCAAAGAAAGGGTAGGC
M305	328673	328695	-							GTTGAGAAGAAGATCCTAGGCC
M306	329472	329496	-							TCCCAGTTACTGCGCGCGATCGTA
M307	329725	329749	-							AGCGTTATAGGTCGTTGGGCGGCC
M308	330786	330807	-							GTGCTATGATTGCCGGAGCCT
M309	351033	351062	-							TCAGCCTTAGTAGAAGTAGGTAGCGGCAC
M310	351660	351693	-							GTAGGACGATGCTGATTGGTTCGAATCCAATGG
M311	361061	361102	-							TGTACTGAGATTGTTCGGGAGACATGGTCCAAGCCCGGTGA
M312	361349	361395	-							GTCGTAACAAGGTAGCCGTAGGGGAACCTGTGGCTGGATTGAATCC
M313	361378	361427	-							CATACCACGGTGGGGTCTTCGACTGGGGTGAAGTCGTAACAAGGTAGCC
M314	362361	362384	-							AACTCAAAGGAATTGACGGGGGC
M315	362672	362692	-							GGTTGAAAGTGAAAGTCGCC

Supplementary Table 4: List of suppliers of chemicals and biochemical

Affimetrix	Affymetrix Inc., Santa Clara, CA, USA
Agilent	Agilent Technologies, Santa Clara, CA, USA
Bio-Rad	Bio-Rad Laboratories, Hercules, CA, USA
Biozym	Biozym Scientific GmbH, Hessisch Oldendorf, Germany
Carl Roth	Carl Roth GMBH & Co, Karlsruhe, Germany
Colgate-Palmolive	Colgate-Palmolive GmbH, Gelsenkirchen, Germany
Duchefa	Duchefa Biochemie B.V., Haarlem, The Netherlands
Epicenter	Epicentre Biotechnologies, Madison, WI, USA
Eurofins MWG Operon	Eurofins MWG Operon, Ebersberg, Germany
GE Healthcare	GE Healthcare Europe GmbH, Freiburg, Germany
Gebrüder Patzer	Gebrüder Patzer GmbH & Co. KG, Sinntal, Germany
Hartmann Analytics	Hartmann Analytic GmbH, Braunschweig, Germany
Illumina	Illumina Inc., San Diego, CA, USA
Kapa Biosystems	Kapa Biosystems Inc., Wilmington, MA, USA
Life Technologies	Life Technologies, Carlsbad, CA, USA
Macrogen	Macrogen Korea, Seoul, Republic of Korea
Metabion	metabion GmbH, Planegg/Steinkirchen, Germany
MP Biomedicals	MP Biochemicals, Santa Ana, CA, USA
NEB	New England Biolabs, Ipswich, MA, USA
PEQLAB	PEQLAB, Erlangen, Germany
Promega	Promega Corporation, Fitchburg, WI, USA
QIAGEN	QIAGEN, Hilden, Germany
Retsch	Retsch GmbH, Haan, Germany
Roche	Roche Diagnostics GmbH, Mannheim, Germany
Sigma-Aldrich	Sigma-Aldrich Corporation, St. Louis, MO, USA
Thermo Scientific	Thermo Fisher Scientific Inc., Waltham, MA, USA
Veolia	Veolia Water Solutions & Technologies, Saint Maurice, France
SMB	Services in Molecular Biology GmbH, Rüdersdorf, Germany
Zymo Research	Zymo Research Corporation, Irvine, CA, USA

Abbreviations

as	Antisense
ATP	Adenosine triphosphate
BLAST	Basic Local Alignment Search Tool
bp	Base pair(s)
CAPS	Cleaved Amplified Polymorphic Sequence
cDNA	Complementary DNA
CDS	Coding sequence
Chr	Chromosome
cpRNP	Chloroplast ribonucleoprotein
DNA	Desoxyribonucleic acid
dNTPs	Desoxy nucleotide triphosphates
DTT	Dithiothreitol
E. coli	*Escherichia coli*
e.g.	*exempli gratia*
EDC	1-ethyl-3-(3-dimethylaminopropyl) carbodiimide
EDTA	Ethylenediaminetetraacetic acid
et al.	*et alii*
GMO	Genetically modified organism
HAT	Half a tetratricopeptide repeat
hcf	High chlorophyll fluorescence
i.e.	*id est*
IR	Inverted repeat
kb	Kilo base pairs
K_d	Dissociation constant
knt	Kilo nucleotides
LB	Lysogeny broth
Ler	Landsberg erecta
miRNA	MicroRNA
MOPS	3-(N-morpholino)propansulfonic acid
MORF	Multiple organellar RNA editing factor
mRNA	Messenger RNA
MS	Murashige and Skoog
mTERF	Mitochondrial transcription termination factor
NDH	NADH dehydrogenase-like
NEP	Nuclear-encoded plastid RNA polymerase
nt	Nucleotides
NUMT	Nuclear mitochondrial DNA

NUPT	Nuclear plastid DNA
OPR	Octatricopeptide repeat
ORF	Open reading frame
PCR	Polymerase chain reaction
PEP	Plastid-encoded plastid RNA polymerase
piRNA	Piwi-interacting RNA
PNPase	Polynucleotide Phosphorylase
PPR	Pentatricopeptide repeat protein
PUF	Pumilio and FBF homology
RACE	Rapid amplifications of cDNA ends
RBP	RNA-binding protein
RIP	RNA editing-Interacting Protein
RNA	Ribonucleic acid
RNase	Ribonuclease
RNA-Seq	RNA-Sequencing
RRM	RNA recognition motif
rRNA	Ribosomal RNA
RT-PCR	Reverse transcription polymerase chain reaction
RuBisCO	Ribulose-1,5-bisphosphate carboxylase/oxygenase
S	Svedberg unit
SD	Standard deviation
SDS	Sodium dodecyl sulfate
siRNA	Small interfering RNA
SNP	Single-nucleotide polymorphism
SSC	Saline sodium citrate
TAL	Transcription activator–like
TAP	Tobacco acid pyrophosphatase
Taq	*Thermus aquaticus*
TBE	Tris-Borate-EDTA
T-DNA	Transfer DNA
TPR	Tetratricopeptide repeat
tRNA	Transfer RNA
UTP	Uridine triphosphate
UTR	Untranslated region
UV	Ultraviolet
v/v	Volume percent
w/v	Mass fraction
Ws	Wassilewskija
WT	Wild-type

Acknowledgements

I would like to express my gratitude to my doctoral advisor Prof. Christian Schmitz-Linneweber for his constant support. He encourages me to pursue my own ideas and after all this years convinced me that working with plants has a lot of positive aspects.

I want to thank Prof. Wolfgang Schuster for being a referee for this thesis after already testing my basic genetics knowledge in the intermediate and final examinations for my diploma in biochemistry.

I am thankful to the designated PPR expert Prof. Ian Small for giving me the opportunity to work in his lab in Western Australia. Beside this six months being scientifically very successful, I had the chance to explore one of the most fascinating places I have ever been to. Thanks to all the Small lab members for the warm welcome. Especially, I thank Kate Howell for introducing me into the lab and sharing her expertise in next-generation sequencing with me. The CRR2 project resulted in a very fruitful collaboration with Peter Kindgren. Peter and Bernard were not only lab neighbors but also showed me some nice places with incredibly expensive beer in Perth.

Michi Tillich was my mentor during the time I did my diploma thesis and the beginning of my PhD studies. He is still an excellent discussion partner.

I really enjoyed working as part of a team with Christiane Kupsch to unravel the function of CP31A and CP29A. I thank the current and past lab members including Ayako, Cori, Jan, Julia, Marie, Marlene, Reimo, Sabrina, Stephie, Yujiao for discussions and the great atmosphere in the lab. I thank Prof. Thomas Börner for the discussions during the group meetings. Gongwei Wang is bringing vital bioinformatics competence into the small RNA team that I am lacking. Thank you for that.

Over the years a couple of diploma, master and bachelor students were keen enough to work under my supervision in the lab. I thank Sandra Gusewski for starting the mitochondrial small RNA project and establishing the small RNA gel blot. Tea and Ella worked on the CP31A project and Arne tried his best to affinity-purify specific RNA-binding proteins. I thank our lab technicians Conny and Jana for their technical support.

I thank my girlfriend Lydia, my family and friends for distraction from work. In the last four years I also had a great time outside the lab, which helped a lot to get over the many disappointments that go in hand with experimental work.

Publications

Loizeau K. *, Qu Y. *, Depp S., Fiechter V., **Ruwe H.**, Lefebvre-Legendre L., Schmitz-Linneweber C., Goldschmidt-Clermont M.: Small RNAs reveal two target sites of the RNA-maturation factor Mbb1 in the chloroplast of *Chlamydomonas*, *Nucleic Acids Res.* 2014

Ruwe H.*, Castandet B.*, Schmitz-Linneweber C., Stern DB.: *Arabidopsis* chloroplast quantitative editotype, *FEBS Lett.* 2013

Kupsch C.*, **Ruwe H.***, Gusewski S., Tillich M., Small I., Schmitz-Linneweber C.: *Arabidopsis* Chloroplast RNA Binding Proteins CP31A and CP29A Associate with Large Transcript Pools and Confer Cold Stress Tolerance by Influencing Multiple Chloroplast RNA Processing Steps, *Plant Cell* 2012

Ruwe H., Schmitz-Linneweber C.: Short non-coding RNA fragments accumulating in chloroplasts: footprints of RNA binding proteins?, *Nucleic Acids Res.* 2012

Ruwe H.*, Kupsch C.*, Teubner M., Schmitz-Linneweber C., The RNA-recognition motif in chloroplasts, *J Plant Physiol.* 2011

* shared first authorship

Selbstständigkeitserklärung

Hiermit erkläre ich, die Dissertation selbstständig und nur unter Verwendung der angegebenen Hilfen und Hilfsmittel angefertigt zu haben. Ich habe mich anderwärts nicht um einen Doktorgrad beworben und besitze einen entsprechenden Doktorgrad nicht. Ich erkläre die Kenntnisnahme der dem Verfahren zugrunde liegenden Promotionsordnung der Mathematisch-Naturwissenschaftlichen Fakultät I der Humboldt-Universität zu Berlin vom 06. Juli 2009.

Berlin, den 15.04.2015 .....

[Hannes Ruwe]